

ED 029 001

By-Cronbach, Lee J.; Snow, Richard E.

Individual Differences in Learning Ability as a Function of Instructional Variables. Final Report.

Stanford Univ., Calif. School of Education.

Spons Agency-Office of Education (DHEW), Washington, D.C. Bureau of Elementary and Secondary Education.

Bureau No-BR-6-1269

Pub Date Mar 69

Contract-OEC-4-6-061269-1217

Note-221p.

EDRS Price MF-\$1.00 HC-\$11.15

Descriptors-*Academic Aptitude, *Individualized Instruction, *Interaction Process Analysis, Learning Theories, Personality, Programed Instruction, Research Design, Research Methodology, Research Needs, Students

Identifiers-Aptitude Treatment Interaction, ATI

This document focuses on how research which investigates the interaction between learning abilities and instructional treatments (Aptitude Treatment Interaction or ATI) should proceed. Previous research related to ATI is evaluated in the context of the ATI premise that characteristics of learners affect their attainment of educational goals (outcomes from treatments). Previous research was found to be inadequate because of weak methodology, inappropriate hypotheses, and lack of replication. Guidelines for future research, introduced throughout the document, encompass design, methodology, and conceptual stages (such as understanding how general ability enters into a pupil's learning). In relation to two ATI goals (different instructional methods for different kinds of students should be employed to achieve the same educational goals, and personality dimensions as well as aptitude should be a criterion for placement rather than for rejection or selection in a program), research evaluation uncovered the following: learning rate is a false issue; general ability is related to learning in conceptual tasks; rote and meaningful instruction may serve different kinds of students; the principles governing the matching of learner to individualized instructional environment are not yet known; and the thinking on personality variables as they relate to instruction is in a primitive state. A 221-item reference list is included. (LP)

ED029001

BR-6-1269
PA-24
OE-BR

Contract No. OEC 4-6-061269-1217
U.S. Office of Education

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE
PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS
STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION
POSITION OR POLICY.

Final Report

Individual Differences in Learning Ability
as a Function of Instructional Variables

Lee J. Cronbach, Director

Richard E. Snow, Assoc. Director

The research reported herein was performed pursuant to a contract with the Office of Education, U. S. Department of Health, Education, and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

SP002635

School of Education
Stanford University
Stanford, California

March 1969

Preliminary Statement

This is a final report on a contract with the Basic Research Branch of the Division of Elementary and Secondary Education of the U. S. Office of Education. It includes a number of tentative statements on theoretical and methodological matters, along with our better established findings from truly completed investigations of specific matters. Because of the tentative nature of some of our most significant statements, we request that investigators proposing to adopt these general ideas or to cite this report communicate with us, so as to take advantage of any corrections we discover and of the further development of our ideas that we anticipate.

To include so much tentative material is uncommon, and a word of historical explanation is in order. The contract, scheduled to run from April 1, 1966 to December 31, 1968, called for a broad exploratory survey of the aptitude-treatment interaction (ATI) problem. The work proceeded much as originally envisioned save that it moved a good deal more slowly, for three reasons. (1) There were conflicting, unforeseeable, and sometimes inescapable demands on the investigators' time. (2) The problems turned out to be much greater than we had anticipated; it was necessary to cut through long-accepted, packed-down -- but false -- premises. (This plowing into hardpan may be the greatest ultimate contribution of the contract effort.) (3) Opportunities for highly significant inquiries, not originally envisioned, presented themselves as the work unfolded.

In the light of these circumstances, we revised our plans in our own mind to allow 12 months of 1969 for reflection and consolidation. In the past, such extensions without additional funds had been approved quite routinely for us and for others. Upon reaching the scheduled termination date of December 31, 1968, we were told that the Office of Education was no longer prepared to allow significant extensions, and that it was incumbent on us to deliver a final report of the work as it stood on that date. In our extensive annual reports, we had indicated the need for a longer time period to arrive at a fully mature report on our problem, and had incorrectly assumed that this was understood by the monitoring officers. Fortunately, we had completed the work specified in the original contract and can report on it here; what is lacking from this report is the information expected from additional studies undertaken, and the more integrated and lucid statement of our emerging conception as to where ATI research should go.

We thank the Bureau of Research for its financial support, and we thank the Stanford Center for Research and Development in Teaching, with which the project has been loosely affiliated, for unnumerable courtesies. We acknowledge the assistance of the project research assistants: Nancy Hamilton Markle, Tamarra Pickford Moeller, Akimichi Omura, and Pearl Roossinck Paulson, and also a large number of other local and distant colleagues, post-doctoral trainees, and graduate students who have contributed to our work and thinking.

L J C
R E S

Stanford University
March 31, 1969

TABLE OF CONTENTS

Preliminary Statement	ii
List of Technical Reports	vi
Abstract	vii
A. A Perspective on the ATI Problem	1
The educational context	1
The methodological context	2
Project aims	6
The social and philosophical context	8
B. Statistical Methods and Designs	14
Conventional analyses	14
Description of interactions as absolute functions	14
Need for absolute statements	15
Nonlinear models	17
Ordinality of interactions	18
Designs	20
Statistical analysis	22
The Neyman-Johnson method	23
The general linear hypothesis	23
C. Learning Rate as a Variable in Educational and Psychological Research	26
Reliability of learning rate	28
Alternative ways of measuring learning	31
Rate scores	31
Limitations of rate measures	34
What to use in place of gain scores	36
The extended course of learning	37
Learning rates at various stages	38
Learning to learn	41
Introduction to the Alvord and Bunderson studies	45
Tuning	46
Correlations among learning measures	49
Correlations for similar tasks	50
Correlations across distinct tasks	51
D. The Structure of Abilities	53
The need for parsimony	53
The hierarchical model	53
Importance of multitrait-multimethod designs	56
The facet model	57
The issue of stability	58
Interbattery research	59
A DAT study	60
Alternative tests of Guilford hypotheses	61
Cluster analyses	62
Facet factor analysis	64
Multidimensional scaling	65
Further work on divergent thinking	66
Analysis of simplex matrices	69

TABLE OF CONTENTS (Continued)

E.	General Ability and its Possible Interactions with Treatment	71
	Correlation of ability with learning	71
	Interactions of ability with programming parameters	79
	Overt response as a variable	80
	"Smooth" vs. "Rough" programs	85
	Miscellaneous FI studies	95
	Meaningfulness of instruction as a source of interactions	106
	Interactions of ability with complex method variables	111
	A study of strategy in verbal learning	121
	A curriculum evaluation study	123
	Teacher expectancy and aptitude "change"	124
F.	Specialized Abilities and Their Possible Interactions with Treatment	125
	Content variables	125
	Discovery or induction in aptitude measure and treatment	143
G.	Interactions in the Personality Domain	149
	Fearfulness	151
	"Structure" as a treatment variable	152
	Task difficulty as a treatment variable	160
	Reinforcement as a treatment variable	163
	Constructive motivation	165
	Other motives	169
	A study of teacher differences	172
H.	Individual Differences in Instruction: Future Prospects	175
	Ways to adapt instruction	175
	Requirements in ATI research	179
	Treatments	181
	Aptitudes	183
	Strategy for investigators and supporting agencies	193
	References	195

List of Technical Reports

- | | | |
|--------------|--|--|
| No. 1 | Cronbach, Lee J. | Year-to-Year Correlations of Mental Tests:
A Review of the Hofstaetter Analysis |
| No. 2 | Cronbach, Lee J. | Intelligence? Creativity? A Parsimonious
Reinterpretation of the Wallach-Kogan Data |
| No. 3 | Snow, Richard E. &
Salomon, Gavriel | Aptitudes and Instructional Media |
| No. 4 | Alvord, Ray W. | Learning and Transfer in a Concept-Attainment
Task: A Study of Individual Differences |
| No. 5 | Hamilton, Nancy R. | Differential Response to Instruction Designed
to Call Upon Spatial and Verbal Aptitudes |
| No. 6 | Cronbach, Lee J. &
Furby, Lita | How Should We Measure Change--Or Should We? |

Abstract

Despite the long history of research on learning and on individual differences, little progress has been made in reaching an integrated understanding of the nature of aptitude or ability to learn. This project sought to assess the present state of knowledge in this area. Specific activities and outcomes of the project, as related to five original objectives, were as follows:

1) A careful review of the large body of relevant literature was completed. One result of the review was the recognition that most of the methodology commonly used in aptitude-treatment interaction (ATI) research was weak and often wholly inappropriate for the uses intended. Suggestions for methodological improvement were formulated. Key points were illustrated using reanalyses of reported data. Another clear finding reaffirmed the substantial predictive value of general mental tests in instructional research. Many studies support the further view that it is possible to establish treatment pairs that have high and low relation to general ability. Studies of narrowly differentiated abilities or those varying programming treatments or content have usually failed to produce ATI. Scattered studies investigating personality and motivation variables as aptitudes were reviewed but no summary conclusions could be justified. Emphasis was placed on the importance of process analyses of instructional tasks as a guide to further research on ATI.

2) The concept of learning rate was reviewed in detail and exposed as a false issue. The significance of the notion of multiple criteria for the learning rate problem was discussed. Both rate measures and level measures were judged inappropriate as representations of learning ability. Approaches to reconceptualization were discussed in terms of multiple regression of outcomes on aptitude information.

3) The meaning of "reliability" of measures of learning rate was also examined. Some alternative estimation procedures were considered. The conclusion was reached that such reliability cannot be determined. Though lower-bound estimates might be obtained by determining maximum predictability using multiple-regression techniques, the problem of deciding whether low validity results from low reliability in a given instance is insoluble.

4) An experiment was designed and conducted to investigate learning-to-learn and the extent of its prediction by mental tests at different stages of the transfer process. The experiment employed concept attainment tasks, asking fifth-grade children to work on seven consecutive problems, the last two involving transfer to a new task family. A control group received only the first, sixth, and seventh problems. Strong learning-to-learn effects were found within problems, between problems, and between task families. Ability tests correlated moderately with concept attainment at each stage; correlations were highest in Problem 5 and lowest in Problems 1 and 6, the first transfer task. The hypothesis of qualitative shifts between stages was examined, by crossvalidating multiple-predictor equations for "early" and "late" learning, and found untenable. Different abilities were not predictive of performance at different stages.

5) Two experiments were conducted to test hypotheses about ATI. In a first study, measures of verbal and figural interpretive abilities were chosen to interact with instructional treatments that made much or little use of graphic-pictorial representation of ideas. Experimental lessons in crystallography were prepared in the two forms and administered to seventh and eighth grade subjects. Aptitude information included measures of spatial orientation, visualization, and verbal comprehension as well as sex and grade. The examination of interactions was conducted by means of the general linear hypothesis. The predictor information did not interact significantly with treatment, either for all cases or for separate grade and sex groups. Evidently the pictorial treatment did not capitalize on specialized spatial talents to a greater degree than the verbal treatment, or vice versa. A number of weak relations, having to do with predictive validity within subgroups, were observed and noted as hypotheses for future research.

A second investigation aimed at comparing a more structured, phonics treatment in beginning reading with more conventional "whole word" instruction. Aptitudes were scales from the Illinois Test of Psycholinguistic Abilities and some experimental measures of short-term memory skills, while criteria were designed to represent both reading achievement and emotional outcomes. A first pilot study, conducted within the beginning two months

of first grade in a local school, showed promising interactions between memory skills and both cognitive and affective criteria. Phonics instruction appeared best for low ability children while whole word treatment served high ability children. A second study, attempting to replicate and improve upon these findings, is being carried out in the same school for the following year.

The report summarizes other experiments bearing on project concerns and reanalyses of previously reported data. General observations on ATI research and educational policy are also included.

LIST OF TABLES AND FIGURES

Table 1, Results of the Maier-Jacobs Experiment on Instructional Programs.	88
Table 2, Results of Reanalysis of Cartwright Data.	91
Table 3, Woodruff Study Treatment Groups	103
Figure 1, Example of Ordinal <u>vs.</u> Disordinal Interaction	19
Figure 2, Learning curves for persons a and b.	33
Figure 3, Learning curves for persons b and c.	33
Figure 4, Nonmetric scaling for postulated order U, C, R, S, T, I, . . .	67
Figure 5, Correlation of concept attainment scores with "general" ability at successive stages (from data of Dunham <u>et al.</u>)	77
Figure 6, Demonstration of reporting from American Institutes of Research <u>Annual Report</u> , 1965.	99
Figure 7, Lorge-Thorndike, Verbal IQ	105
Figure 8, Stallings and Snow Study.	112
Figure 9, Salomon Study.	115
Figure 10, Koran Study.	116
Figure 11, Results of three studies projected onto Melton's model for associative learning.	118
Figure 12, Multiple-regression analysis for Koran study, including video-modeling (VM) and written verbal modeling (WM) treatments	120
Figure 13, Correlations of ability tests with learning trial performance for three treatments. . .	122
Figure 14, Results of Grimes-Allinsmith Study of third grade reading instruction	154
Figure 15, Results from Leith-Bassett study of 10-year olds.	157

A. A Perspective on the ATI Problem

The educational context

The educator devises and applies instructional treatments, continually seeking improved results. One strategy is to seek "the best method of instruction." But pupils differ, and the search for generally superior methods must be supplemented by a search for ways of adapting instruction to the individual.

A good deal of intuitive adaptation, guided by the teacher's experience and impressions of the child, takes place in the classroom. The task of research is to formulate more precisely the ways in which instruction can be varied so as to fit pupil characteristics. Certainly the casual adaptations made by teachers are not the most valid adaptations possible. Indeed, studies of impressionistic judgment, from Binet's earliest examination of students that teachers regarded as intellectually superior, to the latest studies of judgments by professional clinical psychologists, show that biases and errors abound. Very often, the error made is to overdifferentiate, to make fairly radical alterations in the educational program on the basis of limited, transient, or even irrelevant information (Cronbach, 1955).

Homogeneous grouping is an example of ill-regulated adaptation. The conception that learners with good school records should profit from a program that does not suit those who have done badly in school is reasonable enough. But research that asks only "Should the school group students by ability?" is much too limited in conception and has inevitably given conflicting and useless results. What different treatments are to be given to the "fast" and "slow" groups? Obviously, grouping will have scanty effect if the treatment is not varied. And not much benefit is to be expected unless the treatment for each group is patiently redesigned to fit that group properly. But, as many recent writers have suggested, we have not adequately conceptualized the varying kinds of learnings and so are treating as homogeneous groups that need very different kinds of instruction. To group on level of past attainment is to ignore the probable importance of fluid ability and of personal style. Streaming plans are properly condemned as perpetuating social stratification if they are intended merely to simplify the teacher's task and if they have the

effect of setting different kinds of educational goals for pupils with different abilities. If, however, they are designed to move all learners toward essentially the same outcomes -- so far as intellectual and personal development are concerned -- they can overcome stratification.

We know that it is socially indefensible to give some children good education and some poor education. We have captured this in the slogan "equality of educational opportunity." But this too easily degenerates into a Lockean laissez faire which merely invites each child to compete for a place in the system, just as the Declaration of Independence affirms "the right to pursue happiness". Social policy in this century has turned from the passive -- guaranteeing a fair race, but putting all the burden on the individual -- to an active effort to design social conditions that will help everyone run his strongest race. Jensen offers the appropriate slogan for the school: "optimal diversity of educational opportunity." To spell out just what is meant by optimal presents major tasks for the philosopher, the empirical scientist, and the practical educator.

The methodological context

For the empirical scientist, the problem reduces to the search for aptitude-treatment interactions (ATI). To discover and demonstrate these requires a style of research that has only recently become the conscious concern of investigators. Two broad lines of empirical research in behavioral science, the experimental and the correlational, have received extended treatment in writings on methodology and have been illustrated as the standard ways of investigating problems of learning and aptitude. In the past two decades there has gradually emerged a realization that interaction research is a third variety which embraces both the older types of study in a single setting, and so permits investigation of a new kind of question.

Experimental research concerns itself with differences among treatments or conditions; one seeks to establish significant main effects, of the form, say, "Homogeneous grouping plan A works better than heterogeneous grouping plan B." Correlational research concerns itself with the concurrence or covariation of distinct indicators, as in testing such hypotheses as "Good spellers are more successful in learning stenography than poor spellers" or "Independent-minded students are more likely than others to drop out of engineering school." The essential

method is to compare, either by computing a correlation or by comparing means of high and low groups, the standing of persons on two variables.

Interactional ideas are widespread in scientific thinking. An interaction is present when an effect found for one kind of subject or in one kind of setting is not found under other conditions. The possibility of interactions is recognized in the physical scientist's ubiquitous qualifier "Other things being equal..." and in the social scientist's "Can you generalize to other groups (communities, cultures, etc.)" Cronbach (1953), contrasting the method of "correlation between persons" with the conventional correlation between tasks or situations, pointed out that the whole process of seeking laws in science is to somehow partition a grand matrix of organisms and situations so as to obtain sections over which a generalization applies. That is, the task is to group subjects who are similar in their response to some selected range of situations. This kind of theory is especially needed in connection with instruction: What characteristics make instructional situations "similar", in the sense that similar situations are all beneficial for the same kind of pupil? And, in this context, what variables define "similar" learners, i.e., those ready to profit from the same kind of instruction? There is no possibility of theory regarding instruction until learners and learning situations are characterized in reasonably general and comparable terms.

Modern statistical methods for experimental and correlational studies derived from the work of Karl Pearson and his contemporaries. R. A. Fisher, in his series of impressive contributions, advanced methods of both these types, but he also introduced the possibility of systematically testing for interactions. Some technical developments of the 1930's, particularly those arising in Neyman's wing of the statistics department at the University of London, have been almost entirely neglected in behavioral science even though they are highly pertinent to the interaction problem; we shall return to them in due course. As for Fisherian methods of testing interactions (e.g., between species of wheat and effect of fertilizer), these were duly relayed to experimental psychologists, and it is fairly common to see reports of significant interactions of IQ or sex with an experimental variable. These interactions were more often regarded as nuisance than as basic discovery to be interpreted, until they became the focus of special lines of investigation,

such as the Iowa research on anxiety and learning in the 1950's. As for correlational psychologists, they stayed narrowly within the range of studies correlating one test with another or with a criterion, and the factor analytic methods built atop them. The potential significance of interactions was overlooked until the decision-theoretic model forced it to their attention.

In attempting to formulate problems of the tester in decision-theoretic terms, Cronbach and Gleser (1965) realized that an important use of tests was for classification or placement. A classification or placement test cannot be validated by simply correlating pretest with subsequent outcome as in the conventional selection study; what one wants to know is whether the outcome is better under one treatment than another for individuals assigned to each by a proposed decision rule. Classification research existed during World War II, but predictive validity models were used almost exclusively. With the advent of guidance batteries after the war, thought was given to "differential validity", but this was ordinarily stated in what we now recognize as unduly restrictive correlational terms. Theorists studying the classification problem (e.g., Brogden, 1951) recognized the central importance of regression slopes. The decision-theoretic model provided a formal characterization of the problem of validity in placement and classification as the demonstration of aptitude-treatment interactions. One was concerned not only with the existence of ATI but with the benefit to be obtained by using them in decision making, i.e., in allocating persons to alternative instructional (and other) treatments. An ATI exists, in effect, when the regression of outcome under treatment A, upon certain pretreatment information, differs in slope from the regression for the same variables under treatment B. We shall amplify and qualify this statement later.

Substantive interests of the 1950's led to many stirrings of research on interactionist questions. We have mentioned the studies of anxiety as a factor influencing learning, apparently enhancing it under some circumstances and impairing it under others. There was an emergence of an experimental psychology of childhood, where it became apparent that treatment variables rather often interacted with sex. Personality theorists and social psychologists were beginning to acknowledge that the situation in which one subject would function well was not necessarily

the best for another, There were, then, a number of calls upon psychologists to bring interactions squarely into the center of the stage. Cronbach (1957) chided "The Two Disciplines of Psychology" for, on the one hand, regarding individual differences as "error" beclouding experimental effects and, on the other, regarding situational variance as uncontrollable attenuation beclouding the prediction of individual success. He urged a fusion of the two disciplines into one, which would combine correlational and experimental methodology to study interactions. Eysenck⁽¹⁹⁵⁷⁾ insisted that a sound theory of personality or of task performance could not be developed; that a proper theory would have to be a theory of personality and situation.

Despite these stirrings, the movement toward interactionist studies gained speed slowly. Such studies are relatively expensive, the methodology for conducting them is unclear, and the theory that would guide research strategy is little better than speculative. By 1965, the time seemed ripe for major stocktaking. A rather large number of scattered studies had been reported, in various contexts. The concern of educators for adaptation to individual differences was mounting. The problem became live in psychology, as demonstrated by the appearance of the symposium edited by Gagné¹ (1967) on learning and individual differences (based on a 1965 conference), and by many references to person-situation interactions in the Annual Review of Psychology.

Research is necessarily highly specific -- a study of specific subjects exposed to a specific treatment and measured in a specific way. But the main fruit of research is not the microscopic findings of single studies. The main fruit is the conceptualization of nature that is erected by minds reflecting on specific findings. While frequently regarded as almost a by-product of research, it -- rather than the specific findings -- is what guides men in dealing with the world around them.

A second outgrowth of the research effort is a method of investigation, a discipline. Scientists are continually learning to investigate. Just as substantive concepts guide man in dealing with his world, methodological concepts guide investigators in dealing with that world in a more systematic way.

Unfortunately, the logic of research in this new comprehensive discipline has not been clear. At one level, there have been gross failures to capitalize on the data in individual studies. Over and over investigators

have reported demonstrably false conclusions because they have selected inappropriate (though traditional) methods of analysis. There has been far too little realization of the special statistical requirements of interactional studies. We shall return later to some specific examples of the ways in which faulty analysis has wasted money and research effort. Until these faults are remedied, the mix of trustworthy and untrustworthy findings in the literature provides a crumbly edifice upon which no sensible theory can be built. Beyond this level, however, we have found need for a metatheory that will provide perspective for the investigator.

Questions are wrongly posed because of a semantic fallacy. Such statements as "High anxiety goes with erratic behavior" seem to imply functional relations, relations of great importance. But such a statement is ambiguous. From the traditional viewpoint, it means that persons above the mean of a group show more erratic behavior than those below the mean. But no such relative statement is meaningful when we are dealing with interacting phenomena -- and nearly all behavioral and educational phenomena interact.

It is necessary to think through the problem of formulating research questions so that the answers will enlarge understanding rather than obscure reality under false generalizations. So far, in work directly and indirectly connected with this project, we have made some progress toward identifying blind spots in present methods. But this paper represents really only a beginning and much work remains.

Project aims

To put the ATI problem formally:

Assume that a certain set of outcomes from an educational program is desired. Consider any particular instructional treatment. In what manner do the characteristics of learners affect the extent to which they attain the outcomes from each of the treatments that might be considered? Or, considering a particular learner, which treatment is best for him?

Outcomes are plural. Any activity affects many aspects of the person. A method optimal to attain one may have a small effect or even a detrimental effect on another.

This project set out, then, with the aim of providing a survey of the work on ATI, with particular reference to education, so as to clarify

methodological problems and indicate suitable strategies for future investigation, and to clarify the extent to which dependable substantive findings have been established or strongly suggested by the scattered studies made to date. In order to keep the problem as open as possible, "aptitude" has been defined as any characteristic of the individual that increases (or impairs) his probability of success in a given treatment. We have emphatically not confined our interest to what are usually called "aptitude" tests. It seems likely that personality characteristics will have much bearing on a person's response to a given kind of instruction. Moreover, we are unwilling to restrict attention to existing tests that were developed primarily under selection models; tests that predict outcome, and hence are useful in selection, may not be differentially predictive of success under different treatments. If so, new kinds of aptitude must be detected. As for treatment, we again use a broad definition. Variations in the pace or style of instruction are of especial significance to us, but there ought also to emerge, in due time, a general theory covering variables in noninstructional situations as well.

Thus, our concern is a pervasive one aimed at a general style for conceiving and conducting educational research and development in many areas, rather than a subspecialty within any one area. To be sure, the immediate objective of current ATI work is to match specific instructional methods or materials to selected learner characteristics. But more broadly the project is concerned with theory to overarch ideas and proposals as diverse as Holland's ⁽¹⁹⁶⁶⁾ theory of vocational choice, the Pace-Stern work in which the "fit" of personalities to occupational roles or college environments is of interest, Thelen's (1967) emphasis on grouping principles for the improvement of small-group teaching, the branching rules and strategies required in computer-aided instruction, and even the individualized counseling applications of behavior therapy. Further, the ATI view may open important approaches to teacher education, allowing the classification of teachers for alternative training programs aimed at the same ultimate teaching roles as well as the selection of teachers for differentiated roles in school staffs. Teacher characteristics and differences in teaching styles may be seen to function as aptitudes for certain teacher training programs and also as treatment variables in affecting learning of their pupils.

The social and philosophical context

Darwinian theory, especially as interpreted by Spencer and Galton, emphasized "survival of the fittest" as a natural law. So long as mankind was certain to progress, the most highly evolved nation and the most highly capable within that nation became the model. There followed logically Galton's emphasis on selection according to merit and the concept of a single rank-ordering, a g.

Selection was consistent with the Darwinian emphasis on competition -- among species and against natural hazards. Selection by test made the elimination less brutal by changing it to a short, sharp shock.

This was a period of laissez faire; education and social status were goods for which persons would compete, and objective tests only enabled all the likely winners to get into the competition. In the context of 1860, Galton's proposals were liberalizing, as they substituted merit for privilege as the basis for preferment. They opened up "equality of opportunity", but it was an equality of opportunity to compete. The terms of the competition were firmly fixed. It was assumed that the competitive grind of the school system was an adequate basis for finding the best. As Seeley puts it, it was not a society that eliminated slavery but a society where a slave could rise to be an owner of slaves.

Selecting the one type of talent best fitted to survive in the schools was in the end a conservative influence. By reducing the extent to which the schools had to deal with pupils other than the kind they handled best, selection made it less necessary for the schools to invent methods for dealing with other kinds of talent.

Single-rank-order selection is a meritocracy, only a shade less conservative than the aristocratic selection it replaced. It fits only a talent-surplus society. A developed society can use trained persons in large numbers, but has almost no way to use untrained manpower.

The social planner must concern himself not with running a fair competition but with running a talent-development operation that will bring everyone to his highest level of contribution (with due regard to distributional requirements of the society). The complex technical society needs a high percentage of persons in advanced occupations to maintain economic growth and standards of living. Moreover, any disadvantaged segment is a source of social chaos.

The traditional approach of schools has been to select by attainment. Whoever has a good school record to date is favored in the next stage, being admitted to the program that gives more status and perhaps teaches more. Thus educational careers tend to diverge. Early bloomers are favored. Those who do not fit the school as it is are shunted to lower status at each choice point.

The single-rank-order principle loses a large pool of talented persons who are in their way much more excellent (along the lines of dexterity, or leadership, or musical insight, say) than those in the top academic quarter. But multivariate selection programs using tests that are excellently valid for these specialized talents would simply flood the advanced school with prospective failures, so long as the educational methods have evolved to fit only persons high in verbal-academic accomplishments. Identification of the talented is not the basic problem of aptitude testing; the basic problem is to identify those who will do well in, and at the end of, a talent-development program. Something similar can be said regarding training for skilled jobs, where the training is often verbally loaded though the job is not.

Development of aptitude measures and educational methods should be a mutually supporting system, with educational programs designed for the student who does not fit the conventional school and classification procedures designed to choose the right participants for each such program. The old model says: the institution is given, pick persons who fit it. The necessary model says: design enough treatments so that everyone will be able to succeed in one of them. That is a different sort of "equality" entirely.

The successors to Spencer, led by Ward, fostered a Social Darwinism that was essentially a program of environmental improvement. And this involved the idea that some environments are better than others, i.e., the single-rank-order concept was applied to conditions.

Corwin's remarks (1950) put the contrast in views succinctly:

"We are confronted with two interpretations of evolution for social application: The Spencerian, laissez-faire interpretation and the reformist interpretation. Which one was best warranted by the Darwinian doctrine of biological evolution? Inasmuch as Darwin centers his attention upon the struggle for existence among creatures and treats the environment in which this struggle takes place either as relatively inert or as changing in response to factors beyond human control, the answer must undoubtedly be in favor of the Spencerian interpretation.

"Darwin saw all creatures engaged in a struggle for existence, which only those individuals which were best adapted to a particular environment survived to establish new species. From these general premises the laissez-faire conclusion of 'everyone for himself and devil take the hindmost' was perfectly logical if not inevitable.

"The transmutation of Darwinism into a gospel of social reform by Ward in particular required a complete reversal of the formula of adaptation of creatures to environment The formula had to read backward -- instead of the creature being adapted to the environment, the environment had to be adapted to the creature."

The Social Darwinists, however, were not ready to adapt the environment to the particular creature. What was required, they thought, was to create a "best" environment for the species Man. The rank-ordering of environments and of heredities still confuses thought on the heredity-environment problem in psychology. The concept of an heredity-environment interaction, evoked e.g. to explain twin differences and similarities, is still thought of as a ranking of heredities and a ranking of environments, and interaction is often thought of as the reinforcing effect of two pushes in the "good" direction. This view must be supplanted, we argue, by a multivariate conception of environments. That environment optimal for one person is not optimal for another. In a brief paper by Cronbach (1969) (prepared within this project but not issued separately as a Technical Report), this matter is discussed in relation to a lengthy presentation by Arthur Jensen on heredity and environment. Jensen (1969) argues not only that general mental ability is largely inherited, but that blacks are less endowed genetically than whites. Cronbach's comment emphasizes the essential irrelevance of the heritability index, the equal irrelevance of the "enrich the environment" concept, and explains the concept of matching educational environments to the individual. On this last point, Jensen concurs; but his suggestions regarding the nature of the match are quite different from ours.

The interactionist formulation leaves no place for the traditional questions of instructional theory and educational research such as "What is

the best way to teach reading?" The approach is that of the efficiency-expert experimentalist, whether Frederick Taylor or one of his successors. B. F. Skinner (1968) has put the aim of such research succinctly: "We need to find practices which permit all teachers to teach well and under which all students learn as efficiently as their talents permit." Another school of thought, experimentalist in another sense, is the Progressivism that fostered, among other things, the "scientific movement" in education and Deweyan Progressive Education itself (Cronbach and Suppes, 1969). Here again, there was a search for one best strategy, though the strategy allowed for considerable unregulated diversity of treatment.

Educators have adapted their treatments in the past. Progressive education is one example. Another is the change in American colleges, particularly the less elite colleges, in response to the spread of college attendance into the average range of the population. Few modifications, however, have departed significantly from the traditional lecture-course format. Insofar as there have been differentiations, the process of matching students to colleges has been almost entirely disorderly. There is little reason to assume, for example, that it is the person who scores highest on the College Boards who should go to the most selective college. His personal style may be such that he will reach a higher peak if he goes to, say, a small experimental school that is not especially selective.

Attempts to develop compensatory education can be described in the same terms. There is little payoff from a "headstart" program designed to attune children to the same old one-track competition, equalizing footing only at the moment the word 'go' is sounded in the first grade. Compensatory preschool programs have some short-term effects, but before long their pupils are falling behind the pack. Thoughts must turn to an education that is not merely remedial in the narrow "training" sense, but to a true education that employs unique means wherever the child's distinctive development makes traditional methods ineffective for him. The seriousness of this matter depends on the stage of development of a nation. If only a small class of individuals can be put into demanding work, and the economic structure cannot support high-level employment for a large proportion of the talented and educated adults, then selection by one-dimensional ranking is no doubt valid enough. The person who meets the intellectual challenge of school, even the most conventional drill-ridden school, is likely to possess the determination and general

adaptive ability to master other problems. In a developing nation that gives high-school education to less than 5% of its youth, it does not seem at all bizarre to select for academic curricula and the professions the top 1-2% on tests of basic verbal and arithmetical achievements, to put the next 1-2% into vocational and technical training, and to assign those ranking next highest on the tests into clerical training. Better their brains be used in a clerkship than wasted in the role of coolie. For individuals so well qualified academically, it does not seem bizarre that the vocational curriculum in developing countries should use conventional verbal-academic methods. In countries where development is more advanced, so that the demand for well-trained persons begins to match the supply, it becomes important to shift toward differentiated selection, where diverse talents are encouraged, each by its own special educational program. In that economy, the truck mechanic will be a man of dexterity and mechanical comprehension, trained by methods of a concrete sort, and not a failed M.A. who is just below the elite verbally, and is lacking in the concrete aptitudes.

The meritocratic laissez-faire selector and the social-reform experimenter alike missed the point of Darwin's theory. The theory did not posit that generally 'superior' creatures evolve. Spencer may have meant this when he invented the phrase 'survival of the fittest'. Darwin's prejudices may have been similar when he adopted the phrase. But Darwin's theoretical writings are invariably concerned with fitness to survive in a particular environment. If one wants to foster development of a wide range of persons, one must offer a wide range of environments, suited to the optimum development of each person. A social reform that fits environment to the average man or to the present elite is inevitably procrustean, conservative, and self-limiting.

The argument that persons can develop in many ways is not to be confused with blurry values that assume every achievement to be of as much worth as every other. Two kinds of social planning might be considered, and in actuality will blend in various proportions: (1) Distinct roles are allocated to different subgroups, or (2) Common objectives are identified, to be obtained in the greatest degree possible by all persons. Different methods are devised for attaining those objectives. Musical skill is likely to be an objective of the first sort; reading skill an objective of the second.

The only use of differential measurement to date has been in the service of the former type of social planning, based on the thought that different people can learn different things. Our orientation is to the second type of social planning, assuming that information about the learner should help us to plan a way of adapting instruction to him; in Darwinian terms, to provide an environment in which he can thrive.

Adaptation to the individual has been an educational slogan of many schools. Our point is that such adaptation has never been systematic because no one has known the principles that govern the matching of learner and instructional environment.

B. Statistical Methods and Designs

Conventional analyses

Because the study of aptitude-treatment interactions has developed only recently, the methods for identifying such effects are little known, and several bad procedures have been widely used. Before we go on to a technical discussion of advantageous procedures, it will be well to identify the fallacious procedures. These faults will be referred to with monotonous frequency as we attempt to extract conclusions from published work.

Some studies that concern themselves with individual differences report means and aptitude-outcome correlations, without giving standard deviations. But the regression slope, and hence the interaction, is influenced by differences in s.d., and it is reasonable to suppose that treatment outcomes will often have different s.d.'s under different treatments. (Where the investigator has reported s.d.'s, he has often made his interpretation from the correlations rather than the regression slopes, but there the reader can at least calculate slopes for himself and so correct the interpretation.)

Gain scores, calculated by subtracting a pretest from a posttest, are frequently taken as dependent variables. A later section of this report will develop this theme more fully. Basically, the pretest score is an aptitude and should be treated along with other aptitudes. The raw posttest score is the proper dependent variable.

The aptitude variable is frequently "blocked" by dividing the group at the median. This permits 2X2 analysis of variance rather than analysis of homogeneity of regression. Differences in aptitude within the high or low block are ignored. Taking them into account (as the regression test does) would reduce the error term and detect weaker interaction effects as significant. Another procedure, still less powerful in most cases, is to form high, medium, and low aptitude blocks to make a 2x3 analysis of variance. The high and low thirds may differ substantially, but inclusion of the intermediate group can cut the between-groups mean square about in half, bringing it below the point of significance.

Description of interactions as absolute functions.

A basic fault in much behavioral and educational research is the concentration on significance testing to the exclusion of descriptive presentation of results. Even where the effect in a particular study is not significant, a potential contribution is lost if the effects appearing in the sample are not described. The reality of these weak effects may be more

credible if other studies of a similar nature are taken into account. Examination of weak effects also discourages overemphasis on effects within the same study that are not much stronger but that do reach the significance criterion.

The interaction effect is usually thought of as a linear relation of the following type:

$$\hat{Y}_{pt} = \bar{Y}_t + b_{Y_t X} (X_p - \bar{X}_t)$$

This is a linear regression equation; X is an aptitude measure and Y an outcome. It differs from the commonplace regression problem in that the subscript t emphasizes that the means and regression slopes may differ from treatment to treatment. Indeed, a significant interaction effect, by the usual tests, is one in which b_t differs significantly from one treatment to another. A minimum desideratum in describing effects in a study where interaction is possible is to report values of $\bar{Y}$ and b for each treatment, and for each major dimension of aptitude paired with each major dimension of treatment. A more informative way of stating the interaction hypothesis takes the form

$$\hat{Y} = \text{constant} + b_o T + b X + b_t (X \cdot T)$$

Here bX reflects the Y -on- X correspondence common to all treatments (aptitude main effect) and b_t is the increment in slope arising from knowledge of the treatment to which the person is assigned. (With two treatments, T is given value $+1$ or -1 , according to the treatment, and the difference between slopes -- the interaction effect -- equals $2 b_t$. The term $b_o T$ reflects the treatment main effect.)

Need for absolute statements

A good deal of information has been lost through the tendency to conceive of aptitude variables in relative terms. The regression equation is fitted through the sample mean, but it in principle states a functional relation between two absolute measures. By absolute, we refer to a numerical reference frame that is operationally defined, independent of any particular sample. The statement

Expected achievement = $-13 + 6$ (mental age in years)

is such a statement, having an absolute meaning for a particular mental test and a particular achievement test. It is a general working formula that summarizes what we know about the relationship, and tells us what to expect of any learner with mental age (say) 10. Information is lost when a study merely contrasts "High" and "Low" groups (on ability or anxiety or some other variable). In such a study the aptitude means of the groups may not

be reported or may be reported in terms of a nonstandard scale. This precludes future use of whatever relation is discovered.

The importance of absolute statements became clear to us in the course of our reanalysis of the Wallach-Kogan research (Cronbach, 1968; Technical report No. 2). Here several interactions of two kinds of aptitude measure with various dependent variables had appeared to be significant. The authors, and subsequent reviewers, had urged that the relations be "replicated" at other ages. Now it is true that the design could be repeated for groups of younger and older children. But previous discussions made it quite unclear what finding would constitute a confirmation and generalization of the original work. Suppose, in a first study, it is reported that children with IQ above the median profit more from a "discovery" method of teaching, and children below the median profit more from a "didactic" method. Suppose also that the subjects are fourth graders with median MA 10. Now in a third grade where the median is 9, do we expect the median child to show no difference between methods, as did the median fourth grader? Or do we expect him to do better under didactic instruction, as did MA 9 subjects in the fourth grade? Once the problem is posed in this way, we see that the absolute regression function in each grade is required to tell us what our results mean. In the former case the result is evidently social-psychological; the child above the median in a discovery-taught group has the advantage, presumably because he can take a leading part in the class' discovery process. In the latter case the result is evidently individual, and individual mental development determines which method is advantageous.

Methodologically, what is required is that regression equations for various groups be defined relative to the same scale. Preferably, common measures will be used in different studies. Even with this refinement, errors of measurement will cause failures of perfect replication in different samples. Much of this difficulty can be overcome by estimating regression functions for true scores (Madansky, 1959; this matter will be further discussed in a forthcoming publication by Cronbach & Furby, 1969, Technical Report No. 6). Where tests used in one study are too easy or too difficult for the group in the confirmation study, intuitive methods of comparison will probably have to be used to decide whether one function can be regarded as a continuation of the other on some underlying growth scale. Projecting the two measures formally onto a common conversion metric is possible but is an unwarranted complication during exploratory research -- and the study of interactions will remain exploratory for another generation, we suspect.

The preceding paragraph will have caused the reader to think of floor and ceiling effects. These inevitably contravene linear hypotheses, and prevent confirmation, at an extreme level, of a finding well established in the middle range. The linear-regression model that dominates interaction research is also challenged by a number of empirical findings. Especially in studies where anxiety is a pretest variable, it is found that performance under a certain treatment is advantageous for a "middle" group and disadvantageous at the two extremes. Fitting a linear equation will overlook such an interaction.

Nonlinear models. One can inform himself of possible nonlinearity by simply tabulating results separately for high, medium, and low groups. This can be done apart from the basic significance test in a study, since as we noted earlier a 2 x 3 analysis loses power in testing the linear hypothesis. Where nonlinearity is suspected, one can fit a regression equation involving an X^2 term, using a stepwise procedure that calculates the weight for $X^2 \cdot X$ after calculating the weight for X , and tests the significance of the

reduction in the residual. Where there are two predictors, weights would be needed for X_1^2 , X_2^2 , and $X_1 X_2$. One should be hesitant to trust weights for quadratic terms. These are often post hoc. In complicated equations one is fitting a good many regression weights to the data, and very large samples are required to arrive at stable weights (Burket, 1964). Our present recommendation is that the investigator inspect his data for likely nonlinearities and that he mention any impressive ones to his reader, but that he attempt to interpret them only after similar effects have appeared in further studies.

The most basic interaction study, with two treatment groups and one aptitude variable, is in effect fitting a nonlinear function to the data. This was demonstrated in Technical report No. 2, where the Wallach-Kogan data were fitted with functions of the form

$$\hat{Y} = \text{constant} + b_1 X_1 + b_2 X_2 + b_{12} X_1 X_2$$

There was one such function for each dependent variable where a significant interaction was suspected, and the significance of weight b_{12} was one criterion for judging the genuineness of the interaction. In the Wallach-Kogan data there were no treatments; X_1 and X_2 were different aptitude measures. In the elementary experiment with two treatments, X_2 becomes the dummy variable T with values + 1 and - 1 according to the treatment assigned. As Cronbach and Gleser point out (1965), one can think of treatments as

arrayed along a continuum -- e.g., number of frames reinforced in programmed instruction -- and can assign a numerical value to the parameter describing the treatment. Then one is in a good position to describe a function relating outcome to X_1 (aptitude) and X_2 (treatment parameter) jointly, and so to identify the optimum treatment parameter for the individual. This model can be extended to multiple aptitudes and multiple treatment parameters. So far no studies have been discovered where treatments varying parametrically did produce interaction effects, and the continuous model is valuable only as a source of hypotheses.

The hyperbolic paraboloid described in the equation above actually has a more general form which includes two more terms, $b_{11}X_1^2$ and $b_{22}X_2^2$. These terms guard against the possibility that the regression on single variables will not be linear (as in the anxiety research mentioned above). We regard this also as an over-refinement, since it is unlikely that an experimenter will have enough cases to fit more than three weights reliably.

The warnings against overfitting apply with peculiar force to interaction studies because it makes good sense to employ multiple aptitude variables and multiple outcome variables, thus increasing the number of weights fitted. Yet the instructional experiments need to command a good deal of subject time, and hence it is rarely practical to use large numbers of subjects. Sheer empiricism will almost certainly not be a profitable strategy in research on interactions. Only findings that take on theoretical plausibility through their linkages across studies of the same kind and studies of a more basic nature are worthy of concentrated attention.

Ordinality of interactions. One further matter to evaluate is the distinction between disordinal and ordinal interactions. The decision-theoretic model of Cronbach and Gleser, from which interactionist studies took off, places great value on disordinal interactions and none on ordinal interactions. A disordinal interaction is one where the regression lines for treatments intersect, within the range of aptitude of the population in question (see Figure 1). Where that occurs, persons to the left of the crossover point should be assigned to one treatment and those to the right should be assigned to the other, to obtain maximum outcome. Where the interaction is ordinal, the regression functions have different slopes but one is above the other throughout the range, and all persons should be assigned to the corresponding treatment (quota constraints permitting). An ordinal interaction is practically useful however in cases where one treatment is more costly than the other, or for other reasons can be given only to a limited

number of persons. Then the interaction supports a decision rule to assign to the more costly treatment only those persons for whom there is clear chance of benefit.

Our discussion above implies that ordinal interactions should be taken more seriously than previous writings have suggested. An ordinal interaction puts before us the working hypothesis that the regression lines do cross somewhere to the right or left of the range under study. Although Cronbach and Gleser, discussing personnel classification, assumed a fixed population, in psychological research one often studies a subpopulation and is interested in generalizing to other populations lower or higher in the aptitude range. Hence any dependably established ordinal interaction usually implies a disordinal interaction for some other type of subject. A validity-generalization study is required, of course, before the implication can be believed.

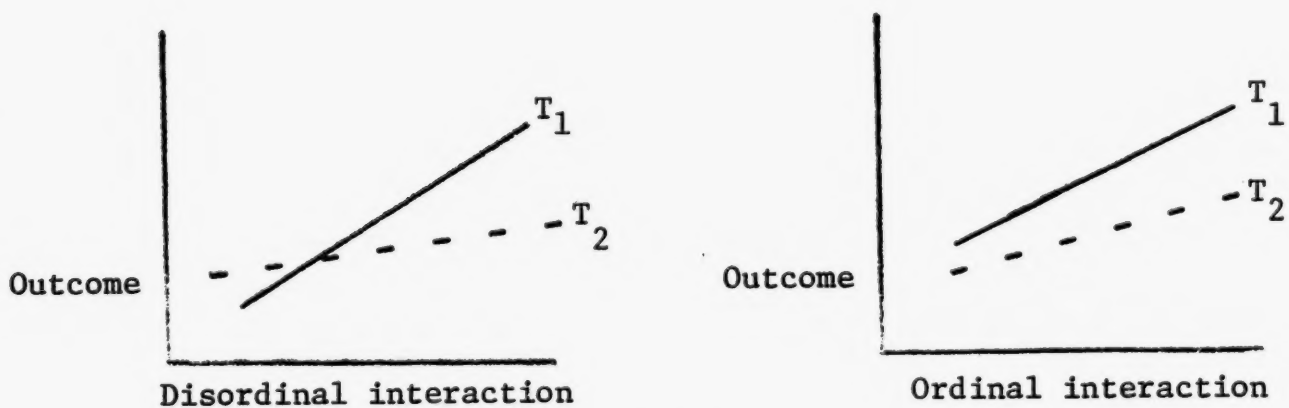


Figure 1

In suggesting that the regression function for one subpopulation can be projected into the range of another subpopulation we should recognize that such a prediction will often be disconfirmed. We spoke, for example, of a regression of achievement on mental age. But mental age 9 implies one thing in a group of adult retardates and another in a group of 9 year olds, as Zeaman and House (1967), among others, have forcibly reminded us. But the solution is not, as these authors suggest, to use IQ as an explanatory and predictive variable. If the regression slope of outcome on MA differs

with age, the correct model is one in which both MA and CA (or MA and IQ, which amounts to the same thing) are used as predictors, in an equation of the form

$$\hat{Y} = \text{constant} + b_{ot}T + b_1(\text{MA}) + b_{1t}(\text{MAXT}) + b_2(\text{CA}) + b_{2t}(\text{CAXT})$$

Here, the weights with no subscript t apply without regard to treatment (i.e., are sheer aptitude effects) while b_{1t} and b_{2t} are increments implying interaction. The regression of Y on true mental age generalizes over ages if adding b_2 and b_{2t} does not produce significant increments in R^2 .

Such a study probably should proceed to weight information in MA and CA jointly by forming the product MA X CA and test whether it improves prediction within and between treatments.

One reason that has been given for minimizing attention to ordinal interactions is that they can be eliminated if either variable is subjected to some particular monotonic transformation. In research where scales are arbitrary, parsimony is served by ignoring effects arising from the scale. The ordinal effect certainly need not be dismissed when the variables are meaningful. The science of genetics makes good use of ordinal interactions, such as that between temperature and number of eye facets developed in fruit flies of different strains. While these scales seem perhaps to be "genuine", expressed in fundamental units, this is scarcely the case. Temperature is an interval scale with respect to linear expansion, but not, surely, with respect to the heat the embryo fruit fly exchanges with its environment or the rate of its biochemical processes. Psychological variables (both aptitudes and outcomes) will become increasingly meaningful as advances in theory explain what underlies a given performance, and superior scales are defined as a result. A deeper reason for taking ordinal interactions seriously is that in decision making one treats the outcome scale as if it were mapped into a nonarbitrary payoff scale. The decision maker cannot decide whether a certain treatment is worth applying unless he can judge the cost and the benefit relative to each other. One transformation of the outcome variable is "right" for that decision maker, and all others are wrong. Also the transformation that makes the regression lines parallel, and wipes out the interaction, may be the wrong one. Educators use tests for decision making, and the scientific ideal of parsimony-in-theory is not always pertinent.

Designs

Nearly all writing on the subject of interactions has assumed a very simple design: an aptitude is measured; persons are assigned at random to

one of two treatments; an outcome is measured; and the interaction is tested for significance. This design, though of basic importance, is inadequate for many significant studies. One may measure more than one aptitude, or more than one outcome. One may have more than two treatments. And assignment is sometimes unavoidably nonrandom.

Extreme-groups designs are often advantageous, if one can sample subjects from a larger pool. One may, for example, choose the highest and lowest fifths of the aptitude distribution, assigning half of each group to each treatment. This is a relatively powerful way of establishing interactions. Interactions are observed in differing regression slopes, and cases widely separated from each other give more information on the slope than cases close together on the predictor scale. It is crucial, however, that when extreme-groups are used to test an interaction for significance, there also be a report of the regression slope. Only that descriptive information tells whether the effect is large enough to be of genuine interest. It is also worth remarking that when cases from the whole range have been put through the treatments there is no advantage in confining analysis to extreme-subgroups.

The extreme-groups design for allocation to treatments can be extended to multivariate aptitude data. With two distinct aptitudes, one will discard cases in the densely packed middle of the bivariate distribution and confine his study to cases around the rim. If there is a strong a priori expectation that the regression slopes will differ in one particular direction, cases on the rim at the opposite ends of that vector will be preferentially retained for the sample exposed to the treatments. Such a commitment to one direction, however, prevents one from discovering that the vector where differences in slope are greatest was not properly located a priori. (The study of Becker, 1967, employed this design.)

The extreme-groups design, whether univariate or multivariate, has one serious weakness; it does not permit one to recognize nonlinear regressions. Ordinarily, however, quite large samples would be required to establish nonlinearity, so that one would not retain the middle cases unless he had a strong hunch that nonlinearity is present.

In an artificial experiment the investigator usually is able to randomize, over all cases or within aptitude levels. In real educational situations, randomization is rare and sometimes out of the question. Cronbach and Gleser, in discussing the validation of tests for placement, envision a sample divided at random between the advanced and slow sections of a course, for the sake of securing data. It is hard to believe that any

school would permit an experiment to assign pupils at random, so that half of the dull go into the fast section, and half the bright pupils are condemned to slow-section boredom. Tests for placement do have to be validated. What to do? It appears that an inference must be made from groups assigned systematically on the basis of aptitude. If a cutting score is established such that all persons above X' go into treatment A, and all below go into B, then one can draw a conclusion from the discontinuity of the within-group regressions (Campbell & Stanley, 1963). The decision must be based on no information except X for this analysis to apply. One would have greater confidence in the result if the school would agree to assign cases at random if they are reasonably near the borderline X' . In studies where assignment is systematic, it is essential that all information used as a basis for assignment be treated as an aptitude in analyzing the data. Otherwise the effect observed may be wrongly interpreted.

Statistical analyses are available that permit highly complex factorial experiments. No very elaborate experiments have been attempted as yet, but it is certainly possible to cross two or three treatment variables in a design and so to isolate components of the interaction.

Statistical analysis

The most rudimentary test for the presence of an aptitude-treatment interaction takes one of two forms: analysis of variance with aptitude entered as a factor, or a test on the hypothesis of uniform regression slope.

Analysis of variance calls for an $m \times n$ design, where m is the number of treatments and n is the number of aptitude levels. As noted earlier, the usual linear hypothesis requires use of only two aptitude levels, high or low. The design can be elaborated to use more than one aptitude or more than one treatment variable. Its chief limitation is loss of power through ignoring within-cell differences in aptitude or through classifying into more than two levels of an aptitude.

The regression test is familiar to many investigators who use it to make preliminary tests of the homogeneity of regression assumption required for analysis of covariance. The finding of significant heterogeneity that novice investigators usually view with distress really signals the possibility of ATI, and should be examined further with that in mind. The logic of the comparison is as follows: if the within-group regressions are the same in the population, each group's mean square deviation around its own regression line will not be significantly less than the deviation around the pooled within-groups regression line. Various versions of the t or F statistic are applicable to testing homogeneity of regression.

The Neyman - Johnson method

A more sophisticated technique was offered a generation ago by Jerzy Neyman and Palmer Johnson (Johnson & Neyman, 1936). Akin to the determination of a confidence interval for a mean, the Neyman - Johnson technique maps, in the aptitude space, a "region of significance" if there is just one aptitude measure, and two treatments, we might be told that there is a region from (say) 8 to 16 on the aptitude scale, outside which there are significant differences in outcome; more specifically, persons above 16 have a significant advantage on one treatment and persons below 8 an advantage on the other. While no doubt there is an observed advantage one way or the other everywhere save at the crossover point (12), the Neyman - Johnson method is conservative, recommending no particular conclusion or decision on the basis of a crossover point that was influenced by sampling error.

A memorandum by Aiken (1968) which grew out of project discussions but was issued independently -- summarizes the theoretical literature on the Neyman - Johnson method and applies it to ATI data. The technique can also be used for multiple aptitudes and multiple groups.

The general linear hypothesis

The general linear hypothesis also dates back to the work of Neyman and Johnson in the 1930's, and is familiar in mathematical statistics (Scheffe, 1959) but we have seen no use of it in the research literature on education and psychology, nor any reference to it in a statistics text or article directed to workers in these fields. The related work of Bottenberg & Ward (1963) is, however, coming increasingly to the attention of educational researchers. (See also Cohen, 1968 and Li, 1964).

This is not the place to attempt a full explanation of the analysis. There is a BMD computer program, but this gives the user little help in interpreting the output. The Hamilton report/ ^(Technical Report No. 5, 1969) gives a moderately detailed account of the procedure and interpretation. Here we shall simply indicate what information can be elicited. Suppose that the treatments have a 2 x 2 design (two levels of two parameters, G and H, crossed to form four treatments). Suppose there is just one aptitude X. There is thus a model of the form

$$Y = b_0 + b_{0G} (G) + b_{0H} (H) + b_{0GH} (GH) \\ + [b_1 + b_{1G} (G) + b_{1H} (H) + b_{1GH} (GH)] X$$

The letters G and H in parentheses stand for dummy variables which take on values of +1 and -1, depending on the treatment to which a subject is assigned. (If $G = +1$ and $H = -1$, $GH = -1$; etc.) There are two types of coefficient, one a set of b_0 's that have to do with effects independent of X, and one a set of b_1 's for regression effects dependent on X. We are primarily interested in the coefficients b_{1G} , b_{1H} , and b_{1GH} , which will be different from zero when there are ATI and not otherwise (save for sampling errors).

The computer model allows us to specify a large number of equations that serve as alternative hypotheses to account for Y. There might be interest, for example, in examining main effects and the GH interaction only; if so, the required equation is

$$Y = b_0 + b_{OG}(G) + b_{OH}(H) + b_{OGH}(GH)$$

One can employ regression methods to "fit the four values of b and to test for significance; this does what analysis of variance does. The computer represents the hypothesis stated above by a series of 1's and 0's: 1 1 1 1 0 0 0. The entry of 1 identifies a term in the original model that is to be fitted; the entry of zero, one that is to be ignored.

To take a different example, consider

$$Y = b_0 + [b_1 + b_{1G}(G)] X$$

This fits a regression equation taking into account not only the overall relationship between X and Y, reflected in b_1 , but the within-treatment regression slope for the two values of G. The hypothesis is encoded, for the computer, as follows: 10001100. The solution will give such values as these: $b_0 = -2.27$; $b_1 = 0.77$; $b_{1G} = 0.28$. This implies that for all cases pooled the regression slope is 0.77; for the treatment where G has the value +1 the slope is 1.05, and 0.49 where $G = -1$. (To write the within-group regression equation, one would need to fit b_{OG} also.)

One can write a very large number of such hypotheses by means of different combinations of 1's and 0's; some hypotheses are more meaningful than others. The computer gives, for each such hypothesis, the Y sum-of-squares accounted for by the prediction, the residual SS, and an F ratio indicating whether the prediction is significantly worse than that given by the basic model 1 1 1 1 1 1 1. In general, to test the significance of any effect, one sets up two hypotheses that differ only in regard to that effect, subtracts the two values of sum-of-squares explained to learn

what predictive information the term in question adds, and divides by the residual SS obtained under the basic model. This F ratio tests the significance of the contrast in question. Several effects may be tested at once, by allowing the contrasted equations to differ in all those respects.

In general, the first desirable step in an interaction study is to test whether there is significant interaction at all; if not, more specific hypotheses should not be tested. To do this, one would specify the hypothesis $1\ 1\ 1\ 1\ 1\ 0\ 0\ 0$ and test whether it gives significantly worse prediction than $1\ 1\ 1\ 1\ 1\ 1\ 1\ 1$, which fits all three interactions. (Note that acceptance of this null hypothesis is required to apply analysis of covariance.) Should there be significant interaction, one might then contrast

$$\begin{array}{cccccccc} 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ & & & & & & & \text{with} \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \end{array}$$

to determine if the interaction of treatment variable G with X is significant.

This method becomes increasingly powerful as one has an increasing number of aptitudes to deal with. The model can be extended to additional predictors simply by adding a further string of weights $b_2, b_{2G}, \dots$ for the second aptitude, and so on. It is not necessary that aptitudes be orthogonalized, but if they are not orthogonalized the interpretation must be more careful. To find that adding a weight b_{2G} does not significantly improve prediction may mean not that the second aptitude fails to interact with G but that it duplicates predictive (interactive) information given by X .

These remarks should be sufficient to suggest how one may proceed to narrow the field of tenable hypotheses and ultimately to pinpoint interactions. The procedure has intriguing complications, but the regular determination of significance levels avoids much of the overinterpretation and confusion that simpler analyses of complex data generate. The procedure has a multivariate extension that can cope with more than one output, but the BMD program for this is still experimental.

C. Learning Rate as a Variable in Educational and Psychological Research

The concept that some people learn faster than others, and that this rate of learning is altered by prior experiences of the person, is intuitively obvious. It has been taken for granted throughout the history of research on learning: the central problem of transfer of training has been defined, since Ebbinghaus, in terms of "savings" in time required to learn, and hence of increases in learning rate. The study of learning rates appears central to research on ATI. Aptitude is, essentially, whatever makes a person ready to learn rapidly (or, outside the educational context, to adapt effectively to his environment). The premise of ATI research is that different instructional conditions call upon different kinds of aptitude, i.e., that a person's "learning rate" will be different in different circumstances. More specifically, the person who learns fast, relative to others, in one condition will be a laggard in another condition.

The concept of rate of learning had always been in the background of educational psychology rather than in the foreground of attention, until the 1960's. During that period, several lines of thought independently drew attention to it.

(1) The new science and mathematics curricula of the period emphasized that in those rapidly changing fields the school was transmitting, not an established culture to be used throughout life, but an ability to comprehend and master findings and concepts as they are created. Put into psychological terms (Cronbach, 1964), the claim was made that the new curricula developed aptitude for learning science (or mathematics, or foreign language, etc.), that the new kinds of study shortened the time a person would require to master a new body of work, or a new language, or whatever. Thus "aptitude" became one of the most important outcomes to be measured in evaluating a curriculum.

(2) The enthusiasts for programmed instruction, taking off from Skinner's 1954 paper, insisted that with proper linear programming every person could master anything; the only differences between persons were in the time required to complete the program. Sad experience tempered these claims, but the discussion of them also produced some new ways of looking at instruction.

A series of distinguished papers by John B. Carroll (1962, 1963, 1965) made time-required-to-learn the central consideration in deciding whether one method of instruction is superior to another.

(3) The attacks of John Anderson (1939) and Woodrow (1946) on the concept of intelligence as ability-to-learn were at last given serious attention. Bloom (1964) endorsed the idea that mental tests reflect only past learning, and that gains in ability from year to year during the school ages are essentially unpredictable. Whether there was one ability to learn or a great many was the subject of research by psychometricians (Gulliksen 1968) and by child psychologists (Stevenson and others 1968). In these studies, then, learning rate was the key variable. As this work proceeded, there was increasing recognition that learning in some kinds of tasks was independent of mental test score and of social or ethnic background. This has been stressed especially by Jensen, who proposes to use these independent kinds of learning as a base for establishing ATI, and teaching the child who had poor performance on the mental test by methods that would capitalize on this independent "learning ability".

(4) The psychometric problem of "measuring change" (and the closely related problem of "overachievement") received renewed attention. To meet the needs of investigators studying personality and attitude change in college, Harris (1963) convened a symposium on the measurement of change. It was not recognized, in that context, that every learning measure is a measure of change, but we have found it important to connect the Harris volume with our own concerns.

We began this project with the premise that -- for any instructional material and procedure -- pupils differ in a hypothetical expected rate of learning. Expected, that is, in the sense that if the pupil were to experience that instruction many times independently, his time-to-criterion would have an average value that we could call his rate for learning under those conditions. One of the most significant outcomes of our work -- if it is accepted following discussion among qualified professionals -- is the recognition that the entire concept of a "learning rate" is a false one, educationally, psychologically, and psychometrically. We are not yet prepared to erect a new conceptual and methodological structure in its place, but we believe that all the lines of inquiry described above must be re-formulated.

Reliability of learning rate

Under our original conception of the problem, we included in the contract a proposal to work on the methodology of measurement of learning rate. As in all measurement theory, the reliability of the measure is a central consideration. Lacking information on reliability, one can never give theoretical meaning to low and moderate relationships; they might arise because the variables compared are psychologically distinct, or they might arise because one of the variables is unreliably measured.

After considerable reflection on the problem and the pursuit of some false leads, we have reached the conclusion that the reliability of a learning-rate score cannot be determined. One can establish a lower bound, essentially by arguing that if the measure enters into a validity coefficient of magnitude r , it must have a reliability coefficient of magnitude r^2 or greater. But this is of little use, when our chief concern is to decide whether a low validity coefficient is due to low reliability.

There are many ways to define reliability. To keep matters simple, let us hold in mind a laboratory task of paired-associate learning for which "parallel forms" exist. Let us take number of errors per trial (with a fixed number of trials) as a score. One could compute something like a reliability coefficient by three methods: (1) Splitting the list of pairs into two halves, and correlating half-scores within each trial. (2) Splitting the series of trials into an odd and an even half, and correlating the two. (3) Administering two lists, and correlating error scores on the two lists. The first two of these are of negligible interest, though they were given considerable attention when the reliability-of-learning-rate problem was under study by such luminaries as Tolman with (Nyswander, 1927), Spence (1932), and R. L. Thorndike (1935).

To split within the trial is to ask only, how adequately have we observed the person's status on this trial? We wish it to be high, since otherwise data points are badly determined. But the data point is perhaps strongly influenced by historical events to which it bears reliable witness; this is not the same as saying that the events are reliable. Person 1 gets off to a flying start; Person 2 becomes seriously confused. This will generate a high split-list reliability, but there is no assurance that the differences would not be reversed when the two independently start to learn another list. Learning is a divergent phenomenon, in the sense that chance

events often have resonating rather than cancelling effects. The subject who confuses two response terms on an early trial will continue to have difficulty for many trials, and his confusion may radiate out into confusion with other response terms. Yet this confusion may be quite fortuitous and uncharacteristic of him. So the score he reaches on trial 4, no matter how precisely it is measured, has an accidental component of unknown size.

The same remarks apply to the method of splitting between trials. Here again, an accidental event can produce consistent consequences. The person who gets confused on some detail on trial one may carry that confusion for so long that he is forever behind his more fortunate competitor. His odd trials and his even trials will tell the same story. But this story of consistence does not imply that we are correctly reporting an inability to learn on his part; on an independent list he might be no more prone to confusion than anyone else.

The difficulty with the list-to-list correlation is of another sort. It is an excellent method of determining the reliability of learning rate if the subjects are in a steady state. That is to say, if their true learning rate is fixed, each list can be thought of as providing a fresh and independent determination; if that is the case, list-to-list consistency will be lowered by any fortuitous confusions that arise. The difficulty is that we can assume neither a steady state nor independence. Learning-to-learn is a well-established phenomenon. People improve their rate from list to list, or more generally, from one learning task to another of the same sort. Many of the problems we want to investigate, in fact, require the measurement of change in learning rate. Even this would be no discredit to the use of list-to-list correlations, if the gains in rate were uniform over persons. But the gains are variable and systematically affected by the learning experiences. Reliability theory requires two "independent" observations. Formally, this requires us to assume, as a minimum, that the person's rank on list B is the same, whether B is learned before or after list A. Unfortunately, fortuitous events on list A do affect performance on list B. Person 1 hits on an effective style of work during list A; whether or not he would do this on any initial experience with the task, he has now done it and carries it forward to make his performance on the next task more efficient also. Person 2 becomes confused. His

specific confusion will not affect list B, which is made up of new stimuli and responses, but his frustration and embarrassment will. Any accidental mishap during learning of list A will tend to lower list B scores. This seems to imply that correlations for successive lists are inevitably too high. But there are other possibilities that reduce list-to-list correlations -- most important, the reduction in variability when a large part of the group has "learned to learn".

The reliability coefficient we need to know is the correlation between two independent learning experiences while the person is in the same state (i.e., before training or other activity alters his expected rate.).

At one time we thought that the coefficient could be estimated indirectly, by taking advantage of the essentially simplicial nature of performance on successive learning tasks. Given a series of lists to learn, we expect the person's learning rate to change. A person's rank changes gradually, in such a way that list-to-list correlations are lower, the more widely separated the two lists are in the series. The matrix of correlations has relatively large values near the diagonal, declining as one moves away from the diagonal. The correlation of one list with the next is presumably attenuated by any change in learning rate that takes place between lists. We considered the possibility of removing this attenuation by a surface-fitting approach. Given the matrix of correlations, one could fit a surface relating r to L_x and L_y (where one could substitute any list number for x and y). Given the surface one could set $x = y$ and obtain an r_{11} for list 1, and $r_{10,10}$ for list 10, etc. This proposal did not stand up under critical examination. While some events attenuate the correlation between adjacent lists, others increase it, as illustrated in the paragraphs above. A fortuitous happening on list 5, which has a lasting effect on technique or morale, raises the correlation of 5 with 6, 5 with 7, etc. Very likely, then, the r_{xx} estimated from the surface is higher than the coefficient we want to know.

By way of practical advice, we stress first the argument to be developed below, that learning rate is the wrong place to focus and that the posttest score should be the center of attention. But then one might wish to ask whether the same posttest level would be reached in an independent learning experience. Seemingly the only answer is to take the shrunken squared multiple correlation between the posttest measure and all available pretest

information (calculated within a group having a common treatment) as a lower bound to the reliability coefficient.

As matters now stand, we are inclined to think that the concept of reliability cannot be properly applied to data in which order effects are suspected. The question we want to answer amounts to asking "What would happen if one could live a segment of his life over many times?" and "How close would that be to what did happen?" Questions like this appear to be beyond empirical reach. If this argument is sound, we cannot properly speak of the reliability of a grade average, learning curve, or any other event. We can speak only of the reliability of observations.

Alternative ways of measuring learning

There are two distinguishable ways of describing the progress of the learner: in terms of level or in terms of rate.

The "level" score takes a reading on the individual at two or more points in time. One example is the learning experiment that presents a number of trials, each of fixed duration, and determines the score (e.g., number of errors, or time on target) for each trial. These are the data that generate the traditional "learning curve," and are also the basis from which change or gain scores are calculated. Use of change scores is unfortunately common in educational experiments, where measurement takes place at only two points ("pre" and "post"), possibly with a third retention measure.

Rate scores. To get a "rate" score, two performance levels are set, and the amount of practice time needed to bring the subject from the first level to the second level is recorded. (Adding a third level permits measurement of rate at a later stage of learning. Any number of successively higher standards may be employed.) In one form, this yields a "trials-to-criterion" measure. Such a criterion as two-consecutive-trials-with-no-errors defines the upper level; it may be assumed, if the task is a strange one, that the learner starts with essentially zero proficiency so that zero is the lower level. More generally, this is called the "common-points" measure. This name is given because the upper and lower performance standards must be the same for all subjects. Yet in many tasks some subjects start at a relatively high level; the lowest level that can be taken as a standard is one high enough that it falls at or above the score on the poorest trial of the ablest subject. Likewise, some subjects may never reach

a perfect score; the upper performance criterion must be reached by all subjects. The common-points method indicates the time required by every subject to achieve the same gain in score. The time some subjects require at the outset to reach the lower standard is not counted against their learning rate.

In retention and transfer studies, the "rate" model generates a savings measure, whereas the "level" model generates a simple measure of achievement on a transfer task or a delayed presentation of the original task.

Although "rate" measures have been prominent in the learning research of laboratory psychologists, they have been rare in educational experiments. The practical difficulties with application of the common-points method in the classroom are obvious; one needs to monitor performance regularly, in order to know when the pupil has attained one of the standards. We proposed originally to reopen the question of the possible usefulness of rate measures in educational work, as well as in some kinds of laboratory research. While it is true that rate and level scores are different ways of processing the same observations, and therefore should resemble each other, they are not interchangeable. Figure 2 illustrates learning curves for persons a and b. Levels 1 and 2 have been set, and times 1 and 2 have been chosen to match the levels for the purposes of the example. Six data points, $a_1, \dots, b_3$ are identified. The broad arrows represent rate measures and the dotted arrows level scores. Now it is evident that the order of the two persons is the same on the two measures (when we recognize that a short broad arrow and a long dotted arrow both report good performance). But note that the rate measures are derived from points a_1, a_2, b_2 , and b_3 ; the level measures are derived from a_1, a_3, b_1 , and b_3 . Hence the two systems of analysis actually select different parts of the data for attention. While Figure 2 showed consistency between the two measures, this is not inevitable. Figure 3 has the same curve for b as appeared in 2; but person c is less superior than a was. Analysis by the rate-of-learning method shows c to be superior, but b is superior by the level measure.

The rate measure is unaffected by transformations in the scoring scale. Any monotone transformation of the scale will generate the same rate measure, since it is stated on the operationally-given time scale.

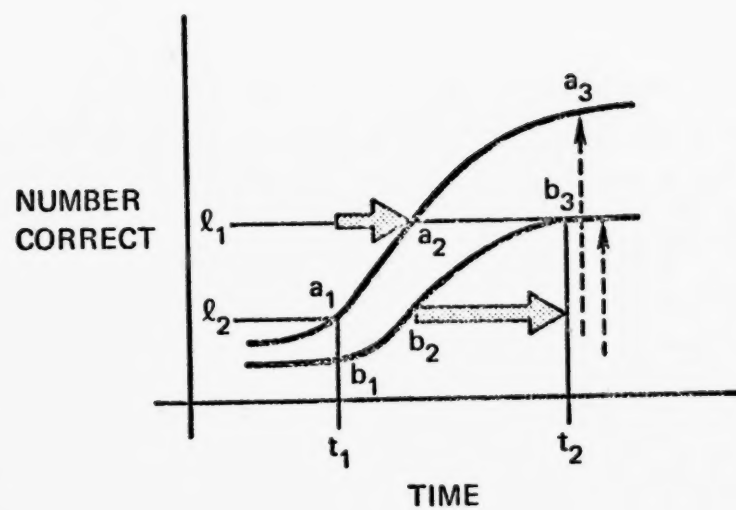


Figure 2. Learning curves for persons a and b.

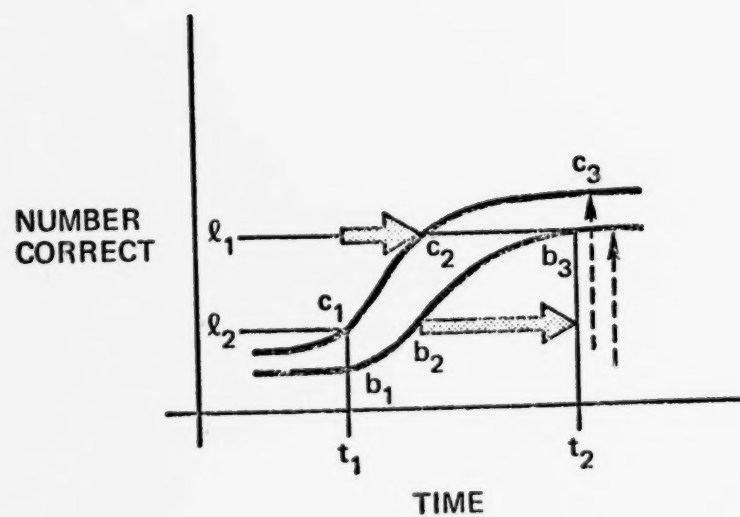


Figure 3. Learning curves for persons b and c.

Also, it is unaffected by ceiling effects that may prevent the superior learner from widening his lead over the rest of the group, or by floor effects that may prevent the person who starts out poorly from showing the full extent of his inferiority. Both measures suffer from a certain arbitrariness in the placement of t_2 , l_1 , and l_2 (it would be usual to take the start of training, and hence it is not arbitrary). Here again, the common-points method seems to have some merit. Ordinarily the lowest and highest standards will be placed as far apart as the group data permit, which gives them a certain objectivity.

We had a further reason for giving attention to the rate measure. If it is true that aptitude is ability to learn, and some curricula are intended to develop aptitude, the obvious way to test such a claim is to compare subsequent learning rates of pupils who had been through the experimental curriculum with those from a control curriculum. Such data are not obtained from the kind of transfer measure, usual in educational experiments, which presents a novel problem and asks how well the person can cope with it. These one-trial insight or application tasks do not allow opportunity to demonstrate learning ability. Transfer studies of the sort required would give a reasonable opportunity to learn the novel material, possibly under instruction. To be sure, one could use either a level score or a rate score during the transfer phase of the experiment, but the rate score speaks directly to the policy question: Does the experimental curriculum speed up subsequent learning to a practically significant degree? (Lawrence, 1954).

As matters appeared at the start of our work, the only disadvantage seen for the rate measure was that it requires a number of successive observations of performance, and cannot be employed in the experiment where there is only a pretest and a posttest.

Limitations of rate measures. As this project proceeded, we found increasing reason to be dissatisfied with rate measures. R. R. Bush & E. P. Lovejoy (1965) in unpublished work where they generated artificial scores under a mathematical learning model, demonstrated that conventional trials-to-criterion scores can be quite unreliable. We are not inclined to believe that this objection by itself is fundamental, save in certain restricted tasks to which the all-or-none model might apply. For more complex performances in which scores progress regularly, smoothing the

learning curves should eliminate most unreliability (save the unreliability arising from accidents that actually impair or facilitate learning, discussed above.) Another critical paper, by Bogartz (1964), is addressed primarily to the procedure in a transfer experiment where a subject who "reaches criterion" is then shifted to another treatment that generates the dependent measures in the experiment. Bogartz, like Bush & Lovejoy, shows that persons with the same true rate reach the criterion at different times, due to whatever chance effects the procedure allows. This in itself is not devastating for the rate measure, since some degree of unreliability is inherent in every measure. Bogartz does, however, make it clear that the unreliability vitiates the intended control in the transfer (or retention) experiments that concern him.

Our own thinking raised further questions that challenge not only the rate measure but the entire line of argument with regard to "savings transfer" and the Carroll model of learning. The essential problem is that performance is multivariate, whereas the measures discussed above conceive of a single outcome to be observed. Even in laboratory research, learning may be demonstrated in several ways which are far from perfectly correlated: reduction of errors, reduction of latencies, and increased resistance to extinction, for example. An analytic view of even the simple paired-associates task indicates that there are several processes of learning, such as discriminating among and becoming familiar with stimulus terms, becoming familiar with and able to produce response terms, and finally linking the two. These could be separately measured and would generate different learning curves. The multivariate nature of educational learning is even more obvious. To follow Professor Carroll in using foreign-language learning as an example: there may be more or less independent development of vocabulary, knowledge of specific patterns, auditory comprehension, pronunciation, and other aspects of successful performance.

To set a standard and determine when the person reaches it, as the rate measure requires, one may (1) set a minimum required level on each of a dozen dimensions, and continue training until every partial criterion is reached, or (2) state a standard on a global, integrative criterion task such as "is able to carry on a fluent conversation on nontechnical matters." Both of these have serious inadequacies. The latter might serve as a crude index of effectiveness of a curriculum explicitly designed to achieve this

outcome and no other. It would fail to recognize the positive virtues of a curriculum that does many things well while neglecting one behavioral element required for fluent conversation so that students are slow to "reach criterion". The difficulty with the more informative compound standard is an administrative one. When the person reaches standard on one subability, are we to assume that no further instructional time is spent on that subability, all effort going into the skills where he is still below standard? It would require an extraordinary efficiency of monitoring and individualization of instruction if we are to make such shifts each time any person reaches a subcriterion. Moreover, it is inconceivable that the subject will not go beyond the standard, or lose some of his skill, as instruction proceeds; the latter can be counteracted by review, but the time-to-compound-criterion measures cannot credit the instruction for any gains beyond the standard that are achieved. The devices sketched here for setting a criterion in multidimensional learning can rarely be used to establish an ordered series of "levels", leading to successive rate measures. This would be possible only if all persons develop mastery of the task in the same way. If one person gains rapidly in pronunciation while another is doing well on grammar but poorly on pronunciation, there is no way to identify a time when they are truly "at a common point" or "in the same stage" of learning.

We decided not to give further consideration to rate measures, then, and searched for other ways of thinking about learning.

What to use in place of gain scores. The investigators who have attempted to use change scores have often arrived at misleading results because of the unsatisfactory psychometric properties of such scores. Proposals by Lord (1963) and McNemar (1958) offered some possible improvement, in suggesting how true gain could be estimated. Another line of discussion led to "residual gain" scores and to something called a "base-free measure of change" (DuBois, 1957, Tucker, Damarin, & Messick, 1966).

Cronbach and Furby (Technical Report #6, 1969) have reexamined these proposals in terms of their psychometric logic, and arrived at two kinds of conclusion.

First, it was shown that one can better estimate true gain and true residual gain by extending the Lord-McNemar approach to bring into the estimation variables in addition to the pretest and posttest. It was shown in the course of this work that the recommendations of several writers regarding the "base-free" measure were incorrect or misleading.

Perhaps a more important outcome of the analysis was a recommendation that change scores should rarely or never be used. It was pointed out that there are four basic purposes for estimating gains. One purpose is to provide a dependent variable in experiments on instruction. There is no need for change measures in this case; the information of greatest importance is elicited by using the posttest(s) as a dependent variable, and using all other information as covariate(s). Second, one might want a measure to serve as a criterion variable in a study where the concern is to decide who learns fastest. This question is essentially meaningless unless persons with ^{the} same true initial status are being compared. When the comparison is made within such a subgroup, the posttest score serves better than any gain measure. The third possibility is to select individuals who deviate from the expected rate of growth, so as to give them special attention or treatment. Again, the analysis essentially calls for a comparison of persons at the same level of true pretest score. The estimated true residual gain (all pretest information partialled out) can be used as an index here, but the posttest score expressed as a partial variable has the same properties. Finally, one might have a construct in mind that can be represented by a difference score. This is unlikely to be the case with differences between pretest and posttests, but differences between aptitudes, personality measures, or measures of performance under different experimental conditions are often treated this way. While an estimate of the true difference score can be made directly by the formulas proposed in the technical report, this involves an arbitrary assumption as to the location of the difference factor in the two-dimensional space defined by the two true scores in question. Very often the most meaningful construct will be better represented by some other factor in the space, and the investigation should be designed so as to leave open the possibility of altering this weight. In sum, then, the paper recommends a number of statistical and psychometric procedures that should be preferred wherever investigators have in the past tried to make use of gain scores.

The extended course of learning

The typical learning experiment treats the data as representing a single process; one may analyze by pretest and posttest measures, or may use a learning curve as a report of gradual change between initial and end points. Increasingly the learning process is being seen as a succession of

developments. Even when there is a single continuous series of practice trials on the same task, many writers now identify distinct "stages" of practice or learning. And in experiments on learning to learn, the learner actually develops into a different, more competent organism as his experience extends in time. Any attempt to study learning ability must form its hypotheses with due regard, then, to temporal changes in the process that is going on.

Learning rate at various stages. Our earlier discussion of the common points method suggests that writers have been fundamentally wrong, when interpreting such work as Fleishman's, to speak about "stages of learning." Fleishman has assumed that during a given trial persons are in the same "stage" of learning or practice. Following this conception, he reached the conclusion that cognitive abilities are more important in early stages of learning, and motor abilities such as speed and coordination more important in later stages.

There is a large body of work, collected in various settings by various techniques, that makes a more general point, of potentially high educational importance. The implied conclusion is that aptitude tests relevant to the early stages of a person's adaptation to a task may not have much to do with his rate of progress during the later stages. This is the Fleishman (1966) finding, in a number of studies of motor skills where the total practice time was quite short (perhaps a two-hour total), and the "early stage" consisted of perhaps the first half-dozen four-minute trials. A very long time span enters the Humphreys finding (1968) that college aptitude tests predict freshman grades well, but have small correlation with grades in later years. Intermediate in duration are the investigations reported by the PSSC and CHEM curriculum studies, (Ferris, 1962) that aptitude tests predict scores during the early part of the year-long course, but have much lower correlations with tests on later units. It is noteworthy that in the educational studies there was no finding of an unconventional test that did predict late-stage scores well, such as Fleishman found for motor abilities. All this work is so critical for thinking about aptitudes that we may digress to make some detailed comments.

The two curriculum studies leave us somewhat dissatisfied for two reasons. One is the absence of any substantial technical report. We need to be assured that the "late-stage" tests were as reliable as those at

earlier stages, and that the same population of pupils was carried through the studies. (If there were dropouts, this alone might account for the finding). The second question is the absence of comparable inquiries conducted in more conventional courses. Until such studies are made, we will not know whether the declining correlations can be attributed to the unusually tight sequential organization claimed for the new curricula or whether the decline is a common phenomenon that happened first to be noticed in the context of some unusual curricula.

What is not certain is that grades in advanced courses are as meaningful as freshman grades. There is undoubtedly a somewhat narrower range of grades, and the smaller classes of the upper years may be conducive to less objective grading. Finally, it seems likely that abler students will take harder courses as seniors than the less able do, thus reducing their chance of high grades.

Despite these reservations, the seeming implication of these studies is that aptitude tests are primarily relevant to the person's success in getting past the initial hurdles of an instructional program. Some learners who start out badly "catch on" to either the logic of the subject matter or the effective style of work, and shift onto a different rate of progress. That is, they learn to learn. And it is possible for morale to deteriorate so that a person is no longer coping as effectively as he did when the situation was fresh. The suggestion of such a finding would be to rely much less on conventional selection plans. To predict that a person will do poorly at the outset is only to advocate patience and special assistance -- assuming that a large number of those who rank low early will rank higher later if they can be kept in the program. The second possibility is that tests other than those hitherto used can predict who will be learning rapidly in the late stages.

It is hard to make a penetrating analysis of what happened in the educational studies, but the Fleishman studies were formally controlled and analytic questions can profitably be asked. It will be recalled that Fleishman and his various coworkers used level measures, such as total number of correct responses made during the time allowed for each trial. These scores were correlated with aptitude tests and it was shown that different aptitudes served as predictors for scores on trials 1-5, 6-10,.... The fact that persons high on cognitive tests generally did well on the early blocks seemed sensible -- understanding directions and patterns of

the task is a cognitive activity, and those who do well on it should get a headstart. The fact that the loading for cognitive tests drops off with the passage of time has been misinterpreted. For the Complex Coordination Test (Fleishman and Hempel, 1954) the percentage of score variance predictable from cognitive tests at different points in practice is approximately as follows:

Elapsed practice time	10	30	50	70	90
Percentage	35	20	10	7	5

Now these data mean that cognitive abilities correlate negatively with the improvement in score between (e.g.) 10 and 50 minutes. This seems paradoxical, especially since we have no reason to search for an explanation in terms of artifact arising from error of measurement. But there is a perfectly clear psychological explanation, which rests on some of Fleishman's own thinking. There is, presumably, a certain amount of cognitive work to be done on a task like this. When the task is analyzed and the basic patterns grasped, the person can perform more efficiently. Hence, whenever he does his cognitive work, his score should rise. The persons who score high on cognitive tests seem to do their cognitive work early in Complex Coordination practice. Presumably the low-scoring subjects also gain the same degree of understanding, more slowly. Hence, they are making substantial gains on some of the later trials (say, around the 20th minute), now achieving what High Cognitives achieved early. Being made by the Low Cognitives, the gains correlate negatively with the aptitude measures. Instead, then, of making the common interpretation that cognitive ability is important in early learning and not in later learning, we offer the hypothesis that cognitive learning is taking place early for some persons and later for others.

At one time we proposed to rescore Fleishman performances in terms of rate instead of level. It is somewhat plausible to regard the progress from one level to another as defining a "stage of learning." While this rescoring might have some value in producing clearer correlation structures, our further analysis of the logic of the problem convinced us that we would not obtain true clarification in that way. Suppose the apparatus allows a person to earn a maximum of 25 points per trial. Then two persons who score 5 points cannot truly be said to be at the same point in their learning, if one of them reaches that score on the first trial and the other reaches it on the fifth trial. They have almost certainly developed different insights into the task, different coordinations, and, in general, entirely different

profiles of subperformances. As Bogartz says (1965), the second person, with more practice, has very likely learned more. Hence, the "common points" method deals with points that are only superficially alike, and to treat them as similar is to discard information. Basically, each person traces a path defined by two coordinates, level and time; more coordinates enter if the measurement of task performance is multivariate. The usual regression model predicts an average performance for persons having similar pre-test scores.

The procedures suggested for estimating true posttest scores are easily extended to the idea of estimating true scores at various points in a learning curve. Thus one can deal with successive measures. This is in effect a kind of smoothing. One may have successive residual gains. That is to say, one can always examine the residual gain between any two points in time. This is likely to be particularly advantageous when one is concerned with retention, since a given retention score will have different meanings depending upon the person's estimated true status at the beginning of the period.

We have stressed the multivariate nature of changes in learning. A method that has recently been developed for handling multivariate data will possibly permit the formation of multidimensional learning curves. The multidimensional scaling methods of Shepard (1962), Kruskal,⁽¹⁹⁶⁴⁾ and others will take a matrix of size n by v by t , that is to say, a matrix of scores of n persons on v variables, each variable being measured at several times t . A scaling procedure reduces the number of variables to two or three. Estimating the person's score on each variable at each time gives points that can be plotted against the v axis. One could actually handle three dimensions without much difficulty, the person's learning track being a twisting line in three-space, on which beads could be mounted to show his position at each successive time. Tracks become much more difficult to visualize if more than three dimensions are retained. To the best of our knowledge, however, in all the work on learning there has never been an attempt to plot even a two-dimensional learning curve, so that the simplest applications of the method are likely to add considerably to our knowledge.

Learning to learn. Any discussion of the validation of aptitude measures must take into account the "learning to learn" (LTL) phenomenon, that is,

the tendency of persons to do considerably better on problems or learning tasks after they have had experience with many problems of the same kind. The learning ability displayed on the first few problems of such a series may not be the most significant indication of the person's ability to perform in an instructional situation or elsewhere where learning will be continued over a long time. As will be seen, very little of the research on aptitudes has taken LTL into account, and one of our purposes was to draw out the implications of this line of possible research for ATI studies.

The existence of LTL as a phenomenon traces back as far as Webb's study in 1917. A comprehensive review of studies during the first half of the century was provided in McGeogh and Irion (1952), where an entire chapter is devoted to the phenomenon. Among the most striking of these early studies is an obscure piece of research by Husband (1947) in which college students worked on a maze and after several months, worked on a second maze. The savings on the second maze were compared with the savings in relearning by a group which encountered the same maze on both occasions. The group having a new maze to learn was less successful, but the difference between the retention and transfer groups vanished after approximately six months. The "forgetting curve" for the retention group had the usual form of dropoff; the similar curve for the transfer group was essentially flat. The implication is that whatever improved learning ability consists of, it is retained to a remarkable degree. Essentially the same finding appears in animal studies conducted by Bunch and his colleagues, (1936, 1938).

LTL was brought back to prominence in psychology by Harlow's well-known experiments reported in 1949 (see also Harlow, 1959). In these studies, monkeys (and in one study, children) worked on long series of discrimination-learning problems and eventually could perform with great efficiency, attaining the correct solution on the first or second trial. This line of research left an unfortunate conceptual heritage, since two fundamentally different types of experiments were mingled together. One is represented by the famous oddity problem, where the correct answer is determined not by the choice of one stimulus as correct, but by the application of the concept that in a group of three stimuli, two of which are alike, whichever differs is to be chosen. A subject who discovers that this rule is in use is able to solve a new problem (new set of stimuli, same solution rule) on the first trial. The other type of problem is one

in which one of two objects is arbitrarily chosen as correct, and the subject cannot possibly achieve better than chance success on the first trial, but can achieve perfect success on the second trial. Recent writers have sometimes called the second type of improvement "discrimination learning set." Certainly in these studies there is evidence of improved ability to learn. In the oddity type of problem, the subject is acquiring a set, but that set is useful only because he is playing a game with an experimenter who agrees not to change the rule. To be sure, the subject is acquiring a generally applicable concept, and thus there is some transferable residue of the experience. But in some ways, this is a much more trivial phenomenon of "learning to read the experimenter's mind." Our primary interest, is in discrimination learning set. In the last two decades there has been considerable work on LTL in children, and much of this is reviewed by Reese (1964). It is evident that in most tasks of this sort ordinarily used in laboratory learning, children do improve with extended practice on successive problems.

Again we may single out one study as being of particularly broad significance. Freibergs and Tulving⁽¹⁹⁶¹⁾ had college students perform on a series of concept-attainment tasks of the sort where a correct answer is defined in terms of the conjunction of selected attributes of the stimuli. In tasks of this sort, it has generally been found that subjects learn much more readily when they encounter a series of instances that are exemplars of the concept (positive instances) than when they encounter a series of non-exemplars of the concept (negative instances). From a logical point of view, however, the two kinds of instances are equally informative and should allow equally early solutions. In the Freibergs-Tulving experiment also, subjects given a series of negative instances encountered considerable difficulty. But the important finding was that the difference between positive-instance and negative-instance subjects nearly vanished by the twelfth concept-attainment problem. Subjects were learning to process information even though they were given no instruction.

It is most regrettable that the many LTL studies have given no attention to individual differences. It would be important to know whether persons superior early in the series of experiences maintain their advantage. If not, most of the research examining whether mental tests correlate with learning rate on an isolated task is irrelevant to the usual educational

situation where a person has successive experiences with similar tasks. It would be valuable also to know whether the person who forms learning sets rapidly is different from the person who is slow in this respect.

The issues were brought into prominence in Ferguson's papers (1954, 1956) where ability to learn was seen as essentially the effective transfer of skills laid down in the past. Ferguson was inclined to think of ability tests as measuring thoroughly familiar performances that had been brought to something of a plateau or limit. In his view, the person successful in learning a new task does so by bringing these basic skills to bear. A very similar idea is seen in the hierarchies of Gagné, where previously established abilities of the sort measured in aptitude tests are thought of as prerequisites for growing new ideas and skills. Reference has already been made to the suggestion that ability to learn is developed during the study of certain new curricula, so that the student who has an advantage early in the course is not particularly advantaged later. The evidence for this is tenuous, and there is no information to indicate whether persons superior in learning the last units of the course could have been identified by aptitude measures at the beginning of the course. The Fleishman research shows the emergence of what appears to be a specific, unpredictable factor, but coordination tasks are so unlike intellectual learning that one would hesitate to generalize.

Before going on to report the available research on individual differences in learning-set formation, it would probably be well to make more specific the significance of this topic for research on ATI.

If general statements about such interactions are to be achieved and used as a basis for policy, they would presumably take some such form as this: "pupils with the following pattern of characteristics learn most rapidly when exposed to instruction of the following type." But this type of generalization is meaningful only if individual differences in learning are fairly consistent throughout an extended instructional program where new lessons are to be learned day after day. If individual differences prove to be stable and predictable, one can capitalize on findings from the experiment in which learning is observed only for a rather short time, often on just one task of a given sort. If individual differences are radically altered during learning-set formation, then the short-term experiments on ATI are likely not to be of practical use, though they may

provide theoretical insights. Even if the persons who learn well after they have become thoroughly familiar with a problem are not the ones who learn well at the outset; this ability to capitalize on experience may be predictable. Certainly the educator would be far more interested in knowing who is going to be capable after he is well into the course, and his approach to lessons has been stabilized, than to identify the student who will be off to a running start.

Two studies of individual differences in LTL have recently become available, one a technical report of this project, and one a doctoral dissertation by Bunderson (1965) at Princeton University. At this point we shall introduce the studies but will present only a part of the results.

Introduction to the Alvord and Bunderson studies. The Alvord (1969) is reported in full in our Technical Report No. 4, and only the main outcomes of the study need to be examined in this summary. Alvord employed concept-attainment tasks of the type used in the Hovland and Wisconsin card-sort experiments, asking fifth-grade children to work on seven consecutive problems. The design was carefully counterbalanced so peculiarities in individual problems would not affect the results. A control group was used which received only the first, sixth, and seventh problems. Alvord had to use quite simple problems, because his naive subjects were unable to cope with complex concepts. Since he did find substantial improvement in ability to handle problems of the type he was presenting, an obvious sequel to his work would be to carry the training further and introduce multiple-cue concepts after the initial LTL has taken place. Alvord was particularly interested in the possibility that a distinction could be made between "learning ability" and "ability to transfer". Transfer was involved at two levels. First, within a family of concepts all pertinent to the same stimulus set, hence with rules involving the same attributes, learning to learn is demonstrated if the subject attains a new rule stated in terms of one of those attributes faster on later problems than on earlier problems. Second, when a new stimulus set is used, whose attributes are different, somewhat more complex transfer is presumably required.

The results showed clear evidence of LTL. The number of trials required to solve the problem dropped from a median of 12 on the first problem to a median of 5 on the fifth problem. Changing to a new set of

stimuli on the sixth problem did not reduce scores. The most surprising finding with regard to central tendency was that the control group nearly equalled the experimental group on the last learning task, even though they had had only two previous problems. If this finding can be trusted, it suggests that two problems separated by a 24-hour interval improve learning ability as much as a series of six problems, five of which are encountered on the first of the two days.

There are many points of similarity between this study and the Bunderson study, and we shall develop the implications of the two side by side. Bunderson worked with Princeton undergraduates, so that he could use a relatively complex kind of problem; he presented a series of 26 problems (but only eight trials for each one). He allowed subjects to record information as they obtained it, whereas Alvord's subjects had to rely on their memories. There was evident learning to learn, though subjects were far from perfect on the last block of problems.

The ordinary view of learning to learn has seen it as an incremental process. It is assumed that the subject gradually becomes aware of the cues to attend to or the ways to direct his attention so that he processes information efficiently. It is quite possible that the gains of the individual subject are much more all-or-none, in the sense that he acquires a particular insight or technique and then holds onto it firmly. This could well be masked if there are many such specific insights that the individual grasps at scattered points in time. Bunderson did identify certain points in the records of his subjects where they made marked changes in strategy, but he did not relate these changes to performance scores. More attention needs to be paid to the individual course of LTL in order to determine whether it is gradual, as the group curves suggest, or sharply discontinuous. Concept-attainment tasks may not be the best experimental vehicle for studies of this kind, because the performance curve of an individual is typically irregular, thanks to transient confusions and "local" insights.

Tuning. Another aspect of the possible discontinuous nature of learning to learn is seen in a number of studies that demonstrated the possibility of what we may call "tuning" the learner. The LTL study invariably leaves the learner to his own devices, and under those circumstances progress toward more effective learning is likely to be slow.

The studies to which we now turn employ a minimal training procedure to suggest to the learner what the nature of the task is and perhaps an effective strategy for coping with it.

One such study is that of Jensen and Rohwer (1965). Children of various ages were asked to do paired-associates learning. Not only was it found that older subjects were able to do considerably better than younger subjects, but that there was a marked social-class difference. There was no such social-class difference in serial rote learning, and it was hypothesized that the advantage of the middle-class children in paired-associate learning came from the fact that they habitually used mediation, that is, interpreted a given pair of words or syllables by some meaningful association. Obviously this would transform the task from rote learning to meaningful learning, and make it much easier. Such a transformation is not possible in the usual serial-learning task, because it is virtually impossible to make up a reasonably meaningful linkage for a long series of unrelated words. To test the hypothesis, Jensen and Rohwer suggested to the lower-class subjects that they attempt to form meaningful connections. This simple bit of advice immediately improved their performance to the point where the class difference disappeared. That is to say, a difference "in learning ability" was erased simply by instructing children to make use of an ability they had but did not consider to be relevant. An extended discussion of this kind of tuning is to be found in a sequence of papers by Flavell and his associates (Flavell, Beach & Chinski, 1966; Corsini, Pitt & Flavell, 1968; Keeney, Cannizzo, & Flavell, 1967; and Moely, Olson, Halwes, & Flavell, 1969). Their research tends to show that at certain stages of development the child simply does not use skills that he possesses; this has been called a "production deficiency". It is shown how rather simple instruction can remove such a deficiency and improve learning substantially.

Another example is a study of discrimination learning by Eimas (1966). This investigator asked children to select a correct answer in a situation where two out of four stimuli had been arbitrarily selected as correct. Naive young subjects had great difficulty with this task because in the early stages of trial-and-error learning they got rewards for different responses. When the investigator did nothing more than explain the design of the task so that the children knew that two of the

answers had been baited with candy, they immediately became highly effective learners. Age was no longer a relevant "aptitude".

Investigators have caused endless trouble for unsuspecting subjects with the "reversal shift" and "nonreversal shift" tasks, because in the standard administration the subjects are given many trials in which red (say) is the correct answer and then without warning the reward shifts to green, or to square. Both of these constitute extinction procedures, but they can also be regarded as partial-reinforcement procedures from the child's point of view, unless he has insight that the investigator is playing an arbitrary game in which the answer key periodically changes. In fact, when the child is warned in advance that the answer key will change unexpectedly at rare intervals, the child has no difficulty in adapting to the shift when it comes. We have not encountered any similar attempt to tune the pupil for efficient learning in the concept-attainment experiment, although both Alvord and Bunderson used a warmup procedure to make sure that the basic rules were clear. (A limited kind of pretraining was provided by Wolff, 1967, and by Osler, 1968.) It is virtually certain that a person could be coached to be very efficient on either the Alvord or Bunderson task, because either of them can be solved by a simple algorithm. These tasks would become trivial if the learner were taught to function effectively; we would need to move on to a more demanding form of concept-attainment to examine any significant reasoning or learning performance.

These comments echo the complaint that David Hawkins (1966) made in a too-little noticed criticism of research on learning. The following extracts will indicate the tenor of his argument.

"Most experimental work in the psychology of learning and teaching has not been very relevant to learning or teaching. A teacher friend of mine put it thus: 'most psychologists,' she said, 'have never really looked at children.' . . .

"to interpret: Let me say something first about the concept of preparation, as when one talks about preparing an experiment. I do not mean the preparation which consists in getting oneself ready, but the preparation of the subject of the experiment, a light-beam, or a colony of paramoecia, or a child, or class-room of children.

"There are many psychological experiments, I know, which require comparatively simple preparation. . . . the preparation involved in pedagogical investigation goes up very sharply with the significance of their results. . . . an experiment which takes a half-hour or a day or a week to prepare is, in general, not

worth doing.

"To call something an independent variable is not to use a name but to claim an achievement. . . . in biology another dimension looms as crucial, that is preparation time. To put a complex system in a prepared state takes time. The good biological experiment has such a long preparation time that husbandry becomes the dominant characteristic of the lab or station; in the short run, at least, its resident prepared species determine its experiments. . . .

"Situations of optimum learning require a great deal of preparation. If we do experiments in learning with only superficial preparation -- instructions, 'training', etc., of short duration -- then the rare things get swamped by statistical noise."

Just as the animal psychologist spends considerable time familiarizing his subject with the laboratory, the existence of food boxes, and other things he needs to know to be a good experimental subject, so the educational psychologist ought to be tuning the subject to the point where he is ready to give an optimum performance. Otherwise, the learning data arise out of whatever tendencies have been left as a residue of his uncontrolled past experiences, now haphazardly activated. The school situation to which we wish to generalize is or ought to be one in which the subject is indeed tuned, since the teacher ought to be showing him how to learn effectively, and will certainly not present tasks where the principal difficulty is to figure out what rules the teacher is following. (To be sure, there are classrooms where the teacher violates these suggestions, but one should be more interested in developing an educational psychology to help the competent teacher than to establish generalizations that will predict what goes on in ill-managed classrooms.) We have, then, gone beyond our earlier suggestion that LTL be provided for in experiments on individual differences in learning; we are now saying that superficial individual differences that result from inadequate understanding of the task or from failure to hit upon an effective strategy should be systematically eliminated by helping the subject to achieve his best style of work within the prescribed instructional technique -- before the experiment proper starts!

Correlation among learning measures

Many investigators have asked whether learning abilities are singular or multiple. Virtually all the research has used controlled, short-term laboratory tasks or classroom adaptations of them. The results of the numerous studies have been contradictory, for a variety of reasons.

Many studies have used gain measures, which are meaningless and likely to be misleading (see above). Other studies have used scores of dubious reliability. This is especially a problem on tasks where insight can occur, in an unpredictable fashion. A third difficulty is that in the absence of a clear conceptualization of learning tasks and the processes they call for, in a particular kind of subject in a particular stage of practice, it is very difficult to generalize. On the whole, we should probably not be much interested in scores made by "untuned" subjects on a brief exposure to a strange task.

Correlations for similar tasks. If learning tasks are quite similar, there are often high correlations among them. Alvord, presenting a series of concept-attainment tasks, found consistently high correlations save for the first task. The need for familiarization (even after elaborate warmup and introductory procedures) implies that pupils cannot demonstrate a stable superiority or inferiority on the initial problem; undoubtedly in some materials several problems would be required before familiarization is complete enough to give stable rates. The fact that Alvord's task-to-task correlations rose notably as pupils became familiar with his task causes us to doubt the usefulness of studies of learning rate that give the pupil no opportunity to learn to learn prior to the critical learning-rate measure. In a few of the studies the subjects had as many as three learning experiences on problems of somewhat the same sort, but most of the data have come from essentially naive subjects. There is a clear need for studies of what we might call asymptotic-learning rate, that is, ability to master a new problem when the type of problem is thoroughly familiar even though the content of the new task is not. This will require considerable thought as to the definition of the class of tasks. On the problems Alvord used, one would expect subjects eventually to reach perfect performance, that is, to process information in such a way that they would achieve the correct answer on the first possible trial. If the tasks remained as simple as those the subjects started with, individual differences in learning rate would vanish. On the other hand, learners who had attained this high degree of efficiency on problems with only two or three attributes might still show substantial differences on problems one or two steps higher in complexity.

As we have said, the intercorrelations of scores after the first were high. Each later problem correlated about 0.60 with its neighbors, and correlated about 0.50 with tasks 3 or 4 places removed from it in the series. This implies that there is a good deal of stability to individual differences even when LT is taking place. Indeed, there are many factors holding down correlations for tasks in this study, so that higher correlations are not to be expected. Alvord's counterbalanced design was so arranged that the sixth problem was different for different pupils, some received a relatively easy problem in the seventh location, etc. Furthermore, whether a problem is hard or easy, confusion can arise; one error in memory may be sufficient to confuse the subject so that he is slow to recover, even though he is generally an able learner. Likewise, it is possible to form a "lucky" hypothesis and, when it is confirmed, to be spared the confusion he would normally experience.

Correlations across distinct tasks. When we turn to the comparison of distinct kinds of learning, the literature is highly contradictory. Since we shall not be able to resolve the contradictions (and, indeed, since virtually never do we have data after tuning or thorough familiarization), we shall do no more than cite representative studies.

Manley (1965) employed three types of concept-attainment tasks: a non-verbal series derived from Goldstein, a card-sort series derived from the Wisconsin and Kendler techniques, and a verbal series derived from Allison. The correlations among scores for tasks in any category were substantial, especially for the Allison tasks where the correlations were around 0.70. The crosscorrelations among kinds of tasks were generally very small. Among 32 correlations of Allison tasks with other concept-attainment tasks, the highest is 0.20 and most are 0.10 or below. Duncanson (1964) employed concept-attainment tasks of the Wisconsin type. Like Manley, he found the tasks to be reasonably correlated with each other, but to have very little in common with paired-associates learning and rote memory.

Stevenson and Odom (1965) used two discrimination-learning tasks and a concept-attainment task. Here, the two types correlated with each other rather substantially (0.40 or better). There was a paired-associates task and anagrams task. (The latter seems not to be a learning task so much as a problem-solving task.) These two tests correlated with each other, but not with the first group of three measures. This rather strongly suggests

a separation of at least two categories of performance, but it does not imply the degree of specificity of the Manley and Duncanson studies. The apparent separation of the two categories, however, is questionable by a further study (Stevenson et al., 1968) where paired-associate learning did correlate with discrimination learning.

D. The Structure of Abilities

The need for parsimony

The general ATI problem calls for relating treatment variables to measurable characteristics of the individual. The number of possible treatment variables is very large, and hence the possible number of combinations to be tested is virtually inexhaustible. The task becomes hopelessly extended if there are also dozens of abilities to be taken into account, all of them equally significant. There has been a tension throughout the modern history of ^{differential} psychology between two schools of thought -- those who wish to concentrate attention on a very limited number of abilities and those who wish to emphasize the diversity of abilities. If the problem cannot be resolved somehow in favor of the former conception, research on ATI must degenerate into a trial-and-error process.

We have not attempted to make a thorough and systematic review of the literature on abilities and their interrelations. But we have given considerable thought to the problem of conceptualization and to the syntheses others have attempted. Without documenting our views, we may indicate the directions in which a solution may lie, and then comment on some more specific investigations we have undertaken.

The hierarchical model. Some form of hierarchical arrangement of abilities now is endorsed by nearly all theorists. This view, sketched out long ago by Burt, Vernon, and Cattell, is now coming more clearly into focus (Vernon, 1965; Horn & Cattell, 1966; Guttman, 1965). At the peak of the system is something variously called *g* or fluid ability or analytic ability, which is now being distinguished from crystallized; verbal analogies using simple words, and quantitative reasoning involving novel relationships, would be intermediate. It is suggested by Cattell that there may be more than one "fluid" ability. He and Eysenck would bring in a separate measure of mental speed (which might give rise to its own hierarchy or be "crossed" with other abilities in defining performance). He would also entertain the possible need to separate off ideational fluency, but this is more controversial. There is also a good deal of evidence that rote memory is separable, and there may well be a separate hierarchy in that domain.

The importance of fluid ability is enhanced by the significant program of work of Witkin (1962) and his colleagues, which shows a pervasive difference in the life styles and intellectual processes of those whom he considers to be "psychologically differentiated" and those less "differentiated". One of his most excellent measures of this differentiation or "field-independence" is the Embedded Figures Test. But this proves to correlate very highly with Block Design and Matrices, and in our opinion is therefore to be regarded as a measure of *g*. Witkin provides persuasive evidence that the fluid cluster of abilities is separable from the more crystallized tests such as Wechsler Information and Arithmetic. Most unfortunately, his major studies have always contrasted high-fluid with low-fluid children; we badly need a study in which fluid and crystallized abilities are both measured; the two are strongly correlated, and in Witkin's reports we cannot separate the effects that are uniquely associated with fluid ability. To begin to understand fluid ability, we need contrasts between high-fluid and low-fluid cases at several levels of crystallized ability.

In this connection, we may parenthetically lament the confusion introduced by the use of IQ as an independent (or, rarely, dependent) variable in many studies. IQ is not a measure of what the pupil can do, but a measure of his standing relative to an age group. In a school where promotion is not automatic, fifth graders having the same IQ are far from alike in level of mental development. Using raw score or mental age may not produce substantial differences in statistical results, but it will lead to far clearer theoretical interpretations. We cannot, however, argue that MA alone is sufficient. There are enough studies indicating that differences between younger and older children with the same MA to argue that CA ought to be kept in view; we would be inclined to keep it as a "second" variable, and ask whether it accounts significantly for any effect after mental age has been partialled out of the data.

Our view of the Witkin studies makes it clear that the borderline between studies of "personality" and "ability" is easily permeated. As a matter of fact, most definitions of general mental ability include stylistic variables. Kagan's (1966) impulsive-reflective dimension is an element in Binet's "power of autocriticism", for example. It seems quite clear that some degree of inhibition of overt responses is

required if one is to bring higher analytic processes to bear; therefore, reflectiveness is a necessary but not sufficient element in g. The amount of inhibition wanted will depend on the problem.)

The argument that fluid ability is a complex repertoire of intellectual strategies, rather than a unified process or a biological parameter, need not be elaborated here. Suffice it to say that complex tasks that require deployment of such strategies seem to provide a useful composite measure even though research may later need to tease the processes apart.

The hierarchical notion is concerned precisely with the teasing apart of abilities that develop somewhat independently. There is argument as to whether abilities that separate out in factor analysis do so because they are necessarily distinct or because the culture now separates them. On the one hand, psychologists are inclined to see mechanical reasoning as an ability that branches off because some children (especially boys) are reinforced for attention to mechanical things and so accumulate a special superiority. On the other, there are scattered indications that specialization of some abilities may be genetically based; one reads, for example, of persons with a genetic anomaly who also have a severe spatial-reasoning disability. Specialization of ability is also to be accounted for by the logic of a system of thought; the person who lags in achieving some key concept in a chain of concepts will lag in developing a whole area of competence. The works of both Piaget and Gagné stress this kind of evolution of competence. The argument is that the patterning of abilities arises from the necessary structure of knowledge itself rather than from either the patterning of environment or of heredity.

Whatever the causation, abilities do differentiate. It is now recognized that there is no single level of abstraction on which psychologists should focus. Sometimes g or some other broad composite serves; sometimes one wishes a profile at the level, say, of verbal, spatial, and numerical abilities; and sometimes one wishes to move down to narrowly specific abilities. Those concerned with programmed instruction are forced to move down to relatively minute abilities ("Possesses the concept of numbers as forming an ordered series", or even, "Knows that eleven comes after ten".) This kind of microanalysis is neither more or less correct than the gross analysis that recognizes only a

few abilities; the question of the size of the bundle into which abilities are tied will depend upon the economies of a particular theoretical or practical proposal. One serious difficulty in arriving at a taxonomy of abilities is that we have so far failed to define what is meant by abilities "at the same level" in the hierarchy. This seems to be more a philosophical problem than a statistical one.

Without attempting to add to such papers as those of Vernon(1961) and Humphreys (1962) on hierarchical systems, we note only that we believe it will be necessary for ATI research to work with rather broad abilities, and employ more specific interpretations only as forced to do so by the data.

Importance of multitrait - multimethod designs. Any time a conclusion is framed in terms of a restricted, specific ability, one is conscious of bothersome alternative hypotheses. Suppose it is found, for example, that the Cubes test enters into an interaction with an important treatment variable. Suppose further that this is thoroughly confirmed, by several experiments. Is one to attribute the effect uniquely to Cubes? or to Spatial Visualization? or to the undifferentiated concept of Spatial ability? or to general fluid ability? Since all these broader constructs do account to some extent for the Cubes score, the matter is left in doubt until a multitrait-multimethod design is brought to bear.

If the investigator's hypothesis is that the significant variable is to be conceptualized as Spatial Visualization (i.e., the ability to visualize the rotation of objects in three-dimensional space), then he must employ two distinct tests of that construct -- perhaps Cubes and DAT Spatial Relations. That is the multimethod requirement. Only if the several tests of the construct show the same interaction is the conceptualization defensible. He must further rule out the broader interpretations by showing that, for example, spatial tests not in the spatial-visualization category do not demonstrate the interaction. With the counterhypotheses stated above, one would want the study to include, as a minimum, a test such as Punched Holes or Minnesota Paper Form Board (rotation in a plane) and a test of nonverbal fluid ability (Figure Series or Embedded Figures, perhaps). Obviously, the more tests are used for each hypothesis and counter hypothesis, the

sounder the conclusion; but there are practical limits. The multitrait-multimethod requirement has not yet been taken seriously in ATI research; we ourselves have come to realize its importance only toward the end of this project, and it is not reflected in the designs of our studies to an adequate degree.

The facet model. The chief modern competitor to the hierarchical notion is the facet model that has largely been developed by Guttman (1966).

One can describe tasks in terms of two, three, or more rubrics -- "facets" or in the sense of the Fisherian experimenter, "factors". Within each rubric there are several "levels", to rely again on the Fisherian term. The array of possible tasks is then the Cartesian product of the levels. Given facets A (a,b,c) and B (1,2,3,4), there are twelve possible cells -- a1, a2,...c4. Within each cell, it may be possible to define a large number of test tasks. Humphreys (1962) has reached the conclusion that the facet model may be considerably more powerful than the hierarchical model. It does not appear that the ultimate solution will be to choose one or the other. If one kind of facet is "content", then there clearly is the possibility of developing a hierarchy within the content area. One should also be able to develop a hierarchy within such a process area as memory. These two hierarchies may well cross into a structure that combines both the hierarchical and the facet model. But this is beyond our present ability to imagine, and certainly the facet model needs to be exploited to the point where we comprehend its possibilities.

So far, the nearest to an exploitation of it appears to be the Guilford search model (1967) -- though mention should also be made of the sketchily reported work of Guttman and Schlesinger (1967). In the Guilford system content, operation, and product constitute three postulated facets. We have made some initial efforts to examine the Guilford data to determine how well they conform to a facet model, and our tentative conclusions will be reviewed below. Personal communication with Professor Guilford, however, indicates that he does not regard the facet model as an empirical hypothesis. If we understand correctly, the famous "box" is no more than a heuristic that suggests cells where factors should be found. The cell factors themselves have no postulated structure; they are to be conceived as parallel "stalks"

in a profile. In particular, two cells that are similar with respect to two out of three facets are not hypothesized to have any closer relationship than cells having no similarity. As a specific example, CSI might or might not be more closely related to CMI than to DMU. Professor Guilford does not insist that there be no relations among cell factors; he acknowledges that an oblique structure might be found, but avoids it because there are no reasonably standard methods of arriving at one best structure for given data. These statements, based on conversations and brief correspondence, may not fully represent Professor Guilford's views. In particular, the fact that his book on the structure of intelligence is organized around the major rubrics such as "cognition" suggests that they do serve as constructs for him. No doubt this matter will be clarified as the question of facet structure is more sharply posed by empirical studies.

The issue of stability. Factor analysis has inquired into the relations among tests given at a single point in time. While one can be interested in momentary states for many reasons, any theory of aptitude will surely have to confine attention to reasonably lasting traits. In practical work, one will be able to adapt instruction to the learner's temporary state. He will vary his tactic if the pupil is bored and restless, or has temporarily gotten rusty on his ability to conjugate être. But any generalized recommendation of a strategy or classification for a pupil and certainly any guidance, should be based on aptitudes that will remain stable over months or years.

There has been some research on the long-term stability of aptitudes measured by the most widely used aptitude batteries, but there has been almost no research on the stability of differences between aptitudes. Suppose there is a hierarchical structure in a certain domain, such that there are, at successive levels, 1, 2, and 9 distinct abilities. Then a stability study should tell us whether the finer differentiations are providing information with long-term meaning. Using methods to be described below, one might find that over a six-month interval third-level information has negligible stability. For the second level, differences such as a-b and b-c might well be confirmed on both occasions. But within category a, the differences found among a₁, a₂, and a₃ on the first testing might not be at all confirmed on the

second. That is, all of the first testing might correlate no higher with a1 of the second testing than with a2 and a3 of the second testing. If so, there is no justification for making the third-level distinction in any decision reaching as much as six months into the future. We believe that this principle will serve as a valuable constraint on the proliferation of factored tests for practical use.

Interbattery research

We now move into discussions of methodology for studying the structure of aptitudes, combining with each discussion the fragmentary results we have. These various lines of work are still under test, and this can be regarded only as an interim statement of progress. We begin with the interbattery studies, a method that bears specifically on the problem of stability raised in the preceding paragraph.

Interbattery factor analysis was developed by Ledyard Tucker (1958), but has been very rarely used. Tucker himself moved on to the more general case of three-mode factor analysis. Most applications of the interbattery method encounter problems because of the ambiguity of the notion of information-common-to-two-batteries.

In^a/doctoral dissertation completed in 1967, Nanda (1967) emphasized a special case where this dilemma does not arise, namely, one where two batteries are regarded as equally good measures of the same information. This would be the case, for example, when the batteries consist of parallel forms of the same test.

In this project we have been interested in the empirical exploitation of the method. Computer programs used by Nanda had to be replaced, due to changes in the available computing equipment. The program now available has somewhat greater flexibility than that originally used. The project had reached the point of exploiting the new programs to produce substantive results of considerable interest, but this work has had to be shelved, due to the decision of the sponsoring agency not to extend the working time of the project. The data in hand include the following:

For the Differential Aptitude Tests Battery, a series of about eight complete intercorrelation matrices from two testings, at intervals of several months.

For the General Aptitude Test Battery, a series of test data from repeated testings in grades 9, 10, 11, and 12 of several thousand students.

For the Wechsler Battery, four different studies in which two forms (e.g., WISC and WAIS) were given to pupils to whom both forms were appropriate (e.g., WISC and WAIS at age 15). If we manage ultimately to complete these analyses without project support, we shall be in a position to recommend redesign of the batteries in two respects: first, with regard to the collapsing of scores that do not give distinctive information stable over an appropriate period; second by recommending extension of certain tests that give weak information on separate dimensions having some stable distinctiveness and perhaps contraction of others that measure some dimensions redundantly. We are inclined to think at this moment that the DAT battery can be reduced to four or five, rather than 8 scores; that the GATB will be very little changed, nearly all of its distinctions standing up under this examination; and that the Wechsler will be radically reorganized. We are certain that the Performance IQ concept will prove indefensible, and anticipate that two or three groups of tests will suffice to carry the "profile" information within the Wechsler. This result would be radically different from that of previous "within-battery" factor analyses.

A DAT study. One piece of work already well along is the reanalysis of one set of DAT data in which Form A and Form M were given to boys in two schools. In one school the tests were given seven months apart, in the other 2 months apart. The data were pooled to give a large sample for this analysis. (The data were supplied by The Psychological Corporation, to whom thanks are expressed.) We shall summarize the results briefly, since we are not prepared to draw conclusions until other batches of data have been processed. This summary will indicate the character of the results to be expected.

There were eight factors, whose eigenvalues were, successively, 3.86, .32, .47, .26, .15, .14, .09, and .07*. This strongly suggests that eight scores are not required, since the information yield of successive factors drops off rapidly. A more specific answer as to the usefulness of the information is to correlate the estimate of the

* These are results for one of the two batteries. Slightly different figures are obtained for the second battery in each case but we shall simplify by giving only one set of values.

true score on each factor made from one set of observed scores with the estimate made from the other battery. This is a reliability coefficient of sorts. For the eight factors the values are .93, .79, .64, .59, .43, .43, .31 and .23 respectively. It seems evident that factors beyond the sixth are useless, at least as the scales are presently constituted. Only the first four factors are well enough measured at present to be worth reporting separately. We therefore rotated (varimax with graphical adjustment) to obtain a simpler structure, with these results (loadings under .25 not shown):

	I	II	III	IV
Verbal	.65	.45	.30	
Numerical	.51	.25	.61	
Abstract	.29	.65	.44	
Spatial		.81		.25
Mechanical		.72		
Clerical			.48	.48
Spelling	.86			
Sentences	.67	.30	.27	
(Reliability)	(.86)	(.85)	(.68)	(.58)

The factors are readily identified as verbal (v:ed), nonverbal (m:k), numerical and abstract reasoning, and clerical speed. The first three could be rotated into an oblique structure. Also, the third and fourth could be rotated to isolate the numerical specific. If these results were confirmed, one would be inclined to recommend simplification of the battery by omission of four scores and extension of the Clerical measure by a second test (perhaps on a second day to reduce the influence of temporary set. If a fifth factor is carried through the rotation, it proves to be a spatial specific, a trifle weaker than those for numerical and clerical.

Alternative tests of Guilford hypotheses

Professor Guilford has constructed tests according to hypotheses suggested by cells of his cube, administered these tests in large batteries, and reported the intercorrelations. He factors the battery and rotates in such a way as to identify tests with the originally postulated factors, to the greatest degree possible. While this is not

open to criticism as a strategy for exploring his system and trying out new test ideas, it is not satisfactory for workers standing "outside" the system. The fit of the data to the hypotheses cannot be well tested by such a method, because small and insignificant differences among correlations/^{are} allowed to locate factors to fit the data in the particular sample. Moreover, the outsider wants to know whether he sacrifices much if he uses a considerably simpler system; even if all of Guilford's postulated factors do exist in some sense, they may not be worthy of much attention if they serve only as "trace elements". Decisions about the Guilford system are of high importance for the future course of work on ATI, since if anything near the 80 abilities he claims to have established must be recognized, future hypotheses will have to be stated in a finely differentiated manner, and elaborate designs to distinguish which abilities are truly relevant to a treatment will be needed.

Incidental to the main work of the project, we have explored in various ways small subsets of the Guilford data. Sooner or later, large scale reprocessing of the Guilford matrices according to alternative hypotheses will be important, but we are not in a position to undertake this work. Our preliminary explorations at least raise some striking questions, and suggest possible methods for later use.

Cluster analyses. Factor analysts rightly contend that merely inspecting tests and grouping those that have high correlations will not necessarily disclose the most useful dimensional structure. Even though that is true, one should be able to use zero-order correlations to check upon postulated structures. Given tests of more or less equal reliability, one expects a higher correlation between tests loaded on the same factors than on tests that "belong to" different factors. If that fails to occur for a particular test, there may be an explanation consistent with the original hypothesis; but if it fails consistently, the factor analysis is failing in its attempt to explain correlations where the two tests allegedly fit cells of the structure that have two facets in common (e.g., same operation, same product). These may be compared with correlations for pairs having just one facet (operation, or product) in common, or none, or all three facets in common. This kind of summary is a more direct test of the hypothesis than the factor analysis itself. Guilford (personal communication)

says
/that he does not expect tests having the same operation in common to correlate especially, unless the tests also share the same value of the other facets. But Guilford must expect generally higher correlations when all three facets are shared.

The Hoepfner-Guilford (1965) study administered 57 tests to ninth graders, most of the tests being measures of divergent thinking. The success of the theory was judged by tallying how often a test was strongly loaded on the relevant factors. Thus a test hypothesized to represent the DFC cell ought to have its highest loadings in the cell; but the tally was made on each facet separately. Thus, it was asked how often DFC, DMU, and other D tests were indeed assigned to D factors of any sort in the factor analysis. We are told that the system was confirmed with the following degrees of success:

"Operation" classification,	74 times out of 81
"Content" classification,	71 81
"Product" classification,	52 81

This result, similar to other reports from the Guilford laboratory, seem to argue for the validity of the system. Even though the "hit rate" for product assignments was relatively low, it must be remembered that there are six kinds of products, so that the rate is well above chance.

We analyzed only the divergent tests in the study, tabulating their correlations with each other. We sorted these into pairs of four kinds: CP, where both tests supposedly reflected the same content and product, Cx, where both tests had the same content by Guilford's system and unlike products; xP, like product and unlike content; and xx, product and content both dissimilar. We then asked, for each kind of test (DMC, for example) what fraction of the correlations of each kind exceeded 0.40. (Actually, we tried several levels of correlation successively, to convince ourselves that the result was not an adventitious effect of the level selected). For DMC the percentages were 33% for CP (i.e., one DMC test with another), 40% for Cx, 7% xP, and 7% xx. Now on the view that tests with more facets in common ought to correlate higher, we would expect high correlations to be most likely in the CP pairings, and rarest in the xx pairings. These figures suggest that the/^{fact of} having similar content does indeed identify functionally

similar tests, and that similar product, as defined by Guilford, has no empirical consequences. When we had tallied the results for 14 different Guilford classifications (DMC being just one), we found that the above figures are almost entirely representative of the whole collection. CP and Cx correlations tended to be equally large, and both strongly exceeded xx. While the xP category was (over all data) intermediate between CP and xP, the effect was a weak one, and reversals not uncommon. Hence we conclude that the "product" classification is of little or no value. Obviously one would want systematically to reprocess as much of the Guilford evidence as possible before finally suggesting how to simplify and clarify the system to salvage its meaningful features.

A number of other small analyses lead us to the conclusion that Guilford factors do not provide a persuasive structure. Sometimes a three-way combination does account for a fraction of the variance beyond that accounted for by one-way and two-way classifications, but is not strong enough to be of practical use. One set of data leads to the estimate that four hours of testing (!) would be required to estimate with reliability 0.70 just the difference between DSC and NSC. If these data are representative, the famous convergent-divergent distinction has extremely questionable practical significance. We have put aside the systematic processing of Guilford data by the correlation-sorting technique because enough has been done to satisfy us that Guilford's scheme is not a suitable point of departure for our development of hypotheses. While the sorting method disconfirms the Guilford system in many particulars, it is not capable of offering constructive counterproposals.

Facet factor analysis. A more powerful scheme for dealing with facet models would simply extract factors representing the several "operations", to determine how much variance that set of hypotheses would account for; similarly for content and product hypotheses. The analysis could be repeated at the second level, extracting the content x operation factors, for example, from the first level residuals. Finally, one could test whether there is indeed any appreciable residual variance accounted for by "cell" factors. This would be a far more parsimonious approach than Guilford's, and would make it much clearer where an elaborate scheme like his has value (if anywhere). The basic

model for such a facet breakdown was stated by Guttman in 1958, and has simply not been exploited. Details of the factor analytic process have not been worked out, though we have a tentative flow diagram. Not having reached the point of testing the procedure and identifying its weak points, we are not prepared to report it here.

One difficulty is that for adequate interpretation an "orthogonal" design is required, with all cells represented by the same number of tests. Only very small subsets of the Guilford matrices conform to this requirement. We believe, however that larger submatrices can be selected where slight departure from orthogonality will becloud the interpretation, but to a degree that is tolerable.

Multidimensional scaling. Nonmetric scaling techniques, developed by Lingoes and Guttman (Lingoes, 1965), Kruskal (1964), and others, permit the identification of a wider variety of order or patterning relations than do clustering techniques or factor analytic methods, and may offer more parsimonious representations as well. Guttman (1965) provided one example of this by reinterpreting the original Thurstone PMA studies. Guttman arrives at a distinction between tests measuring "analytic ability" and those measuring "achievement" and shows some of Thurstone's "primary" factors to be subordinate within the achievement category.

An advantage of such techniques, in addition to their use as general surveying devices, is in examining the facet and ordering characteristics implied, if not specifically hypothesized, in Guilford's model. Using Kruskal's program for exploratory purposes, we have reprocessed two correlation matrices from Guilford's laboratory (Hoepfner and Guilford, 1965; Hoepfner, Guilford, and Merrifield, 1964). Scaling the complete matrices or the subset including only recommended factor tests yielded neither a simplified reclassification, comparable to Guttman's reworking of Thurstone's data, nor any clustering in keeping

with Guilford's model. Guilford's "content" classification does seem to produce more coherent clustering than does his "product" classification. Second, we processed selected slices from the cube, to test hypotheses about the ordering of product factors. Figure 4 shows two dimensional plots for product tests, obtained separately for three content categories using the Kruskal program. The view that tests drawn in order from the units (U), classes (C), relations (R), systems (S), transformations (T), and implications (I) cells form a hierarchy or ordered array fits divergent semantic tests (DM_)--but not divergent figural (DF_) or divergent symbolic (DS_) tests. Note in part a) of the figure that an ordered series may be traced from C through R, S, and T. The U tests form a cluster roughly equidistant from other points on the curve, perhaps suggesting a structure more complex than simple hierarchy.

These and other forays into selected portions of Guilford matrices yield interesting implications for further work, which the current project has been unable to pursue. Our experience with the nonmetric scaling methodology, however, suggests that such further work might be extremely profitable. A range of structural hypotheses can be tested, using data already available in reports from Guilford's laboratory, though it is likely that many other hypotheses will require the construction and administration of test batteries not customarily found in Guilford's past research.

Concurring with Harris (1967), we
/ strongly recommend a vigorous program of reanalysis of the Guilford data according to schemes other than the hypothesis-determined simple structure he employs. His writings are obviously influencing many investigators, and if his proposals are unduly complex -- as our preliminary work suggests -- a great deal of subsequent research effort will be misguided by them. To establish firmly the techniques of facet factor analysis and scaling analysis as procedures readily available for psychometric research would be no mean by-product.

Further work on divergent thinking

When this program of work was beginning, one of the most vigorous ideas afoot was the conception of "creativity" tests, derived from Thurstone's fluency tests and other sources, and elaborated by Guilford, Torrance, and others. Just as we report above that the distinction fades almost to nothing in certain Guilford data, so writings

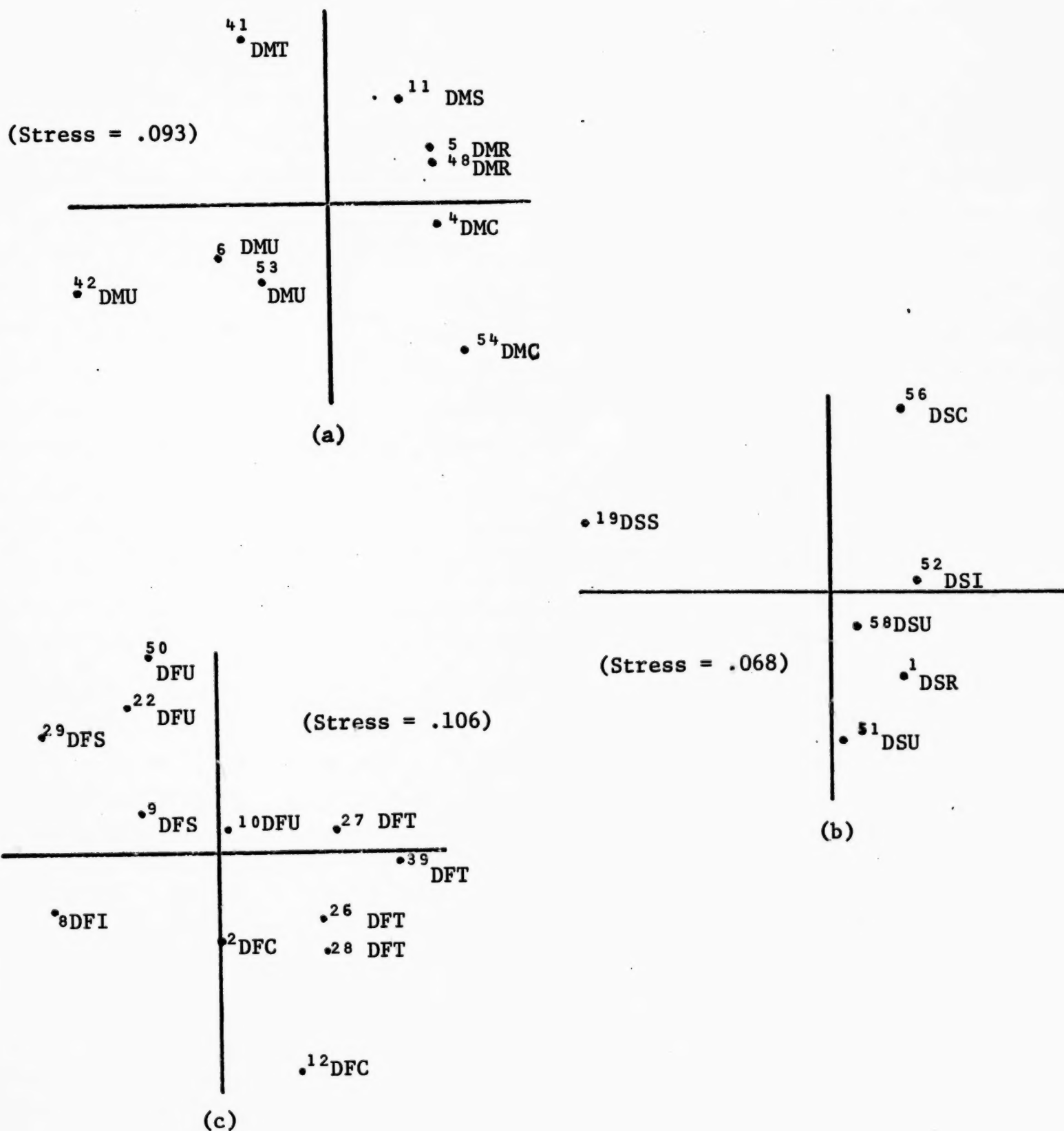


Figure 4. Nonmetric scaling for postulated order U, C, R, S, T, I using a) 9 divergent semantic (DM_) tests, b) 6 divergent symbolic tests (DS_), and c) 12 divergent figural tests (DF_). Stress values (Kruskal's Goodness of Fit Measure) less than .10 represent fair to good fit using two dimensions. Numerals identify points corresponding to numbered variables in Hoepfner and Guilford (1965).

of the early 1960's had indicated that the Torrance and Getzels-Jackson tests did not pull away very clearly from commonplace measures of general ability. In 1965, however, Wallach and Kogan (1965) had published a book claiming that some modified techniques of administering fluency tests produced scores that were independent of conventional mental tests and that had strikingly different correlates in child behavior. They went on to propose thirteen pages of "implications and applications for education", many of which urge different instructional procedures for different children, according to their standing on the two kinds of test. We therefore took the study quite seriously as proposing ATI hypotheses requiring evaluation.

There is no doubt that the F (for fluency) measures are uncorrelated with the conventional A (for achievement) measures. But as we dug into the remainder of the study we found the data quite inadequate to support the important conclusion that high-F children were a talented group, distinctive in many ways. The claim, that the fluency tests measured something intellectually significant, rested on reports of statistically significant relations with a great variety of measures of classroom behavior, problem solving and personality. The report had many puzzling aspects. The significant relations had inconsistent patterns and made less sense to us than to Wallach and Kogan.

We therefore obtained the original data and designed a reanalysis, more powerful than the original one. Our reworking of the data produced a greater proportion of significant relations, and relations that were psychologically more consistent. But few significant relations involving the fluency variable survived the reanalysis. The results were more suggestive of vigorous activity among those scoring high on F than of any intellectual superiority. We found nothing to support the identification of this variable with "creativity", and little to support the elaborate conceptualization of ability-personality relations offered by Wallach and Kogan on the basis of the original pseudo-significant results. The final outcome of our work was a reinterpretation of F as capable of identifying maladjusted high achievers, but not as having direct implication for instruction or ATI research.

This work (prepared by Cronbach, 1968, as Technical Report Number 2) has now been published, with particular emphasis on demonstrating the

methodology that concentrates on large and powerful relations while minimizing statistical noise.

Analysis of simplex matrices

The term simplex has been applied to matrices in which correlations near the diagonal (adjacent trials, for example) are high, and correlations drop off in a certain systematic manner as they go further from the diagonal. These are commonly approximated in learning data, for example in the Fleishman studies, as described earlier in this report.

One study by Hofstaetter (1954) attracted our attention because it claimed to have established, by factor analysis, that there are three qualitatively distinct factors in intellectual development in early childhood. This work had been cited uncritically by many leading authorities on intellectual development. (Indeed, it had been selected for republication in a collection of "model research studies" until our report was made.) Since it appeared that the conclusions were unsupportable from a psychometric point of view, Cronbach (1967,

Technical Report Number 1) prepared a paper demonstrating the fallacy of the work. Hofstaetter had applied a simple-structure factor analysis to data from Bayley's successive mental tests of a sample of young children. His analysis was technically correct, but given an unjustified interpretation.

The factors extracted from a simplex are determined mathematically, in a way that divides the range of the data into a limited number of segments. Which ages are identified with a factor will depend strongly on the starting and ending points of the whole series of measures. Since Hofstaetter's series ended at age 6, he found a single stable factor from age 4 upward. When we treated simplex data for older children this Hofstaetter factor broke into several segments corresponding to the range of the new group. This demonstration was repeated for various samples, until it was obvious that the findings depend almost entirely on the design of the study and hence do not reflect nature.

This work has broader implications. We have indicated earlier that interest was aroused by Fleishman's finding of a specific factor in each psychomotor test he studied. Such a finding is not wholly artifactual, but it does derive heavily from the simplex form of his matrices. In some studies, he has found both an early simplex

(decreasing over trials) and a late specific (increasing in influence over trials); this is even more clearly a consequence of applying Thurstonian methods to a simplex. Had he tried to extract a third, intermediate specific, he could no doubt have done so, arriving at a pattern like Hofstaetter's. In general, we would recommend that Thurstonian methods not be applied to simplicial data; neither the mathematical model of the simplex nor our conception of the way growth data and learning data are generated is consistent with a simple-structure interpretation.

E. General Ability and Its Possible Interactions with Treatment

Correlation of ability with learning

A necessary preliminary to a review of studies of interactions is a survey -- however partial -- of correlations of abilities with outcomes under a single treatment. We are basically interested in the correlation of pretest measures with attainment following instruction, and the well-known predictive validity of aptitude tests might seem to make any question about such correlations pointless. But the ability of tests to predict school success is often explained away by noting that these tests serve as measures of past achievement, and ^{perhaps} predict successfully only because the pupils who have achieved well in the past have a headstart toward the next point of measurement.

Among psychologists it is widely believed that the relation between general ability and learning ability is an open question. Woodrow's attack on those who had defined intelligence as "ability to learn" was slow to make an impact, but gradually psychologists thinking about learning rates came to accept his view. As one example, consider this statement from an investigator whose students have carried out the majority of recent major studies on differences in ability to learn: "We have as yet no clear results on the relation between aptitude tests and performance on learning tasks" (Gulliksen, 1968, p. 798). Likewise, Stolorow (1966, p. 138) seems to take Woodrow at face value. Now there is room for greater clarity, but it will be overwhelmingly clear as we proceed with our summary of literature that some kind of general mental ability almost invariably correlates with attainment in school-like learning. Correlations range from 0.20 to 0.70. This is true even where the content is essentially new, so that no question of measuring gain arises. It is true for a good many laboratory tasks as well as for instructional learning.

There are some exceptions to this generalization. In particular, laboratory rote learning does not often correlate with general tests. But performance on connected instructional materials correlates with tests even under mechanical (so-called rote) instruction. The widespread tendency among specialists to acquiesce in Woodrow's negative generalization arises from many causes: use of rote tasks, use of unreliable learning measures and overly complex factor analytic breakdowns, use of gain scores and other illegitimate forms of dependent variable, concentration on retardates where special problems of interpretation arise, etc.

If, on the basis of the evidence to be presented, it is agreed that there is a common tendency for treatments to relate to g or v:ed or overall past achievement, this does not rule out the possibility of worthwhile interactions. It may be possible to design treatments that depend much less on general ability than is usually the case, or to design alternative treatments that depend on general ability but have different special-ability correlations.

In stating a firm generalization, we warn against using the word "intelligence" in interpreting it. We know that children who do well on tests do well in much learning. Individual differences probably can be substantially altered on some tests or learning measures by "tuning", and we know nothing about what this will do to the correlations. So an hereditarian interpretation is speculative and, we suspect, wrong. It will also be noted that we have not sharply identified the responsible ability as crystallized or fluid. Solid data to support statements as to which ability relates to learning do not exist.

To arrive ultimately at interactions one must find some treatments for which tests -- conventional or not -- do predict and other treatments for which they do not. Hence it is valuable to review what is known about ability-outcome correlations even under single treatments. The most powerful studies are those emanating from ETS and Princeton University over the past ten years. They have had in common the use of many diverse ability tests, on fairly large samples of subjects, and the collection of a number of measures of success in learning under short-term laboratory conditions.

Allison (1960) related code-learning, concept attainment, and psychomotor tasks to aptitude tests. This is one of many studies where an ambitious factor analysis seems to have clouded essentially simple results. In studies with numerous variables, tests are usually short, and correlations are small. Factor patterns then are likely to rest on small amounts of variance. Let us ignore the factor analysis. There are 481 simple correlations (on 315 subjects) between basic learning rate measures and reference measures. Sixty-two of the correlations equalled or exceeded 0.30, but 55 of these are attributable to just three of the 13 learning tasks, and five more are attributable to a fourth. Moreover, if we look at the three predictable learning measures (CIC Plotting, Verbal Concept Formation II, and Spatial Concept Formation II), we find that their correlations spread over almost all the tests in the reference group. Some tests predict better than others (Number Series, Letter Sets, Vocabulary, and General Classification, among others). It appears

Learning was predictable by tests. There is little patterning to the correlations of individual tests, however, and among the best predictors are the highly general mental tests. The implication is that, in this series of concept-attainment tasks, general mental ability is an equally good predictor for early and late problems. Scores on Problem 1 were a bit more difficult to predict than scores on later problems. Any test has nearly the same correlations across problems 2 to 7. Lorge-Thorndike and Stanford scores correlate with learning in the range 0.20 to 0.30. Among the tests from factorial studies, one test was apparently irrelevant, and the others have patterns like those for the standardized tests but at a lower level (partly explained by their lower reliabilities). The one test in this group whose correlations exceed 0.30 on many problems is Hidden Patterns, which can easily be interpreted as a measure of fluid or nonverbal reasoning ability. These data seem to deny that different tested abilities account for performance at different stages in practice, or that transfer requires different abilities than initial learning. In this respect, the pattern of correlations is quite unlike that of the Fleishman studies. A multiple-correlation approach was used to obtain a better idea of the extent to which learning was predictable. Six variables, selected from the original set, were used as predictors, with the following results:

	Problem	1	2	3	4	5	6	7
All cases ($N = 147$)		.32	.38	.40	.40	.46	.35	.36
90 randomly selected cases		.49	.41	.52	.41	.49	.42	.43
57 holdout cases		.02	.17	.08	.30	.35	.12	.09

One would have concluded from either of the first two rows that the correlations are nearly uniform across problems; from the crossvalidation results in the last row we learn that the other multiple correlations are seriously inflated, and in a way that distorts the pattern of data. Worse than that, the manipulation has introduced noise into the system so that the cross-validation R 's are lower than the zero order r 's for several of the variables. The difficulty appears to have arisen from the tendency of one of the factor tests to pick up negative weights in the 90-case sample. After full consideration of these complex results, we believe that the best statement about the predictability of concept-attainment scores is given by the zero-order correlations for Lorge-Thorndike total IQ, to wit:

.29 .35 .32 .31 .40 .29 .29

that we account for virtually all the sizeable correlations if we simply say that these three learning measures are strongly dependent on general ability. When we look also at correlations in the range 0.20 to 0.30 we find two or three more learning measures that can be predicted, almost always by the tests most like the conventional mental tests.

The factor analysis did establish a rote-learning factor that could be predicted from rote-learning tests. A highly-specialized factor representing some aspect of psychomotor learning also splintered off.

On the whole, then, the most important finding appears to be that general ability is related to learning in conceptual tasks. This finding has reappeared repeatedly.

We turn next to the Manley (1965) study, since he used some of the Allison conceptual tasks. He did indeed find these to be predictable on a modest level by seven of his sixteen tests; with one exception these are measures of reasoning or fluid ability, and the best single predictor is Logical Reasoning. In this study, however, card-sort concept attainment tasks and another kind of nonverbal concept attainment were not predictable.

In the Alvord (1969) study previously introduced, aptitude tests from the French and Guilford sets were employed as predictors, along with scores from the Lorge-Thorndike intelligence test and the Stanford achievement test. The single predictor of performance on a concept-attainment problem was performance on the immediately preceding problem; combining several preceding problems gives a still better prediction. The squared multiple correlations for predicting scores on problems 3 to 7 from scores on preceding problems fell in the range 0.40 to 0.48 (implying correlations in the range of 0.60-0.70). Taking ability scores into account improved the prediction to some extent, but this increase is evidently largely due to chance. Apparently, then, tested ability plays its most important role in getting the student off to a good start. Among students who are performing equally well on the fourth or fifth problem, tested abilities can tell us little about who will improve his standing on the sixth problem. Conversely, the student who gets off to a good start on the first few problems, despite the fact that his ability tests gave a poor prognosis, is likely to maintain his standing. Once he has become facile in solving such problems, even if his initial success was a result of happy accident, successive experiences are likely to maintain this skill.

These statements are supported by the observation that the matrix of residual interproblem correlations, after removal of the variance predicted from aptitudes, is very nearly in the form of a simplex. That is, the data suggest that on each trial some individuals are making gains in ability that help their work consistently thereafter. These gains are, so far as we can judge, fortuitous and unpredictable, in the way that flashes of insight are. It would also be consistent with the data to think of adventitious losses entering for certain persons and depressing their performance for some time thereafter. Either gains or losses or a mixture could generate the simplex pattern.

The finding that general ability predicts equally at all stages, but that no ability measure adds much to prediction beyond that given by previous problems, is not in agreement with the report of Dunham, Guilford, and Hoepfner (1966). They conducted a somewhat similar study with high-school students. The concept-attainment task used was not in the general pattern of the Wisconsin cards, as Alvord's was, and they provided little opportunity for learning to learn. They gave three problems, one on each day, and collected scores at various points in the subject's work on each problem. Hence their analysis is for stages within a problem whereas Alvord's scores came from successive independent problems. To illustrate their procedure, consider the figural task given on the second day. S guessed which of four classes each successive figure belonged to -- one class, for example, consisted of figures with two intersecting lines. He could eventually infer the defining rule for each class in turn. There was also a symbolic task (first day) and a semantic task (third day). A large number of Guilford tests selected primarily to emphasize the "class" type of "product" were administered. The usual Guilfordian factor analysis was carried out, extracting 16 principal axes (some of them accounting for little variance); rotation was designed to discover the postulated structure. This structure seems extremely tenuous; as nearly as can be judged without laborious calculation, unrotated factors after the first have negligible correlations with the concept-learning scores. Because Alvord's data also identify no worthwhile difference among kinds of aptitudes, we have here asked only how the concept-learning scores relate to the first principal axis of the Dunham data. Our method was a crude one: we selected the five tests with highest loadings on this factor. These tests (Figure Class Inclusion, Letter Grouping, Verbal Classification, etc.) are novel, but seem clearly to involve reasoning and are not too dissimilar to more conventional mental-test tasks; a straight vocab-

ulary test ranked only slightly behind these tests in factor loading. We then determined the median correlation of the five tests with the concept-formation score at each stage. Figure 5 shows a rising correlation, up to a value of 0.25 - 0.45, depending on the problem. The correlations are somewhat lower than Alvord's, perhaps because of unreliability of the short tests in this study. We do not find convincing evidence that the three kinds of problem relate to different abilities or that scores on the third problem correlate differently from the first (save for an earlier rise in r). The data stand as firm support for the idea that general ability does correlate with this kind of learning (or problem solving).

The Bunderson study, previously introduced, gives a similar result. Multiple correlations for predicting his most basic scores rise steadily from 0.45 on the first block of problems to 0.60 on the last two blocks. (This is contrary to the Alvord finding, and to the Fleishman finding of decreasing correlations for cognitive tests.) We are not given zero-order correlations among block scores, nor are we given beta weights, so that interpretation of the difference between studies is hazardous. Bunderson does give covariances of factor scores with block scores. Among reasoning tests, two factors show their greatest loadings for blocks 2 and 3, and two others peak during blocks 3 to 5. Three memory factors peak on 2 or 3, and one gives a flat function from 3 to 6. Two miscellaneous factors show declining patterns. At face value, this argues that different kinds of ability are important at different stages of learning. But what is baffling is the observation that the multiple correlations increase steadily, even at a time (blocks 4-6) where the zero-order correlations are tending to decrease. Only one factor shows much increase in r from block 4 to block 6; one suspects that as in Alvord's case the rise in multiple R is fortuitous, and a consequence of the entry of negative weights into the formula.

As to the finding that there are different relevant abilities at different points, one needs to note first that the factors are generally correlated. The three factors having the greatest relationship to problem-solving are verbal reasoning, general reasoning, and a novel kind of measure, memory-for-chunking. The first two rise a bit in influence after the first two blocks, and memory for chunking loses some influence on the last two blocks. But the tendency of different factors to be influential at different points is a weak one. We strongly suspect that a simple sum of the tests would predict almost as well as a regression equation (when crossvalidated),

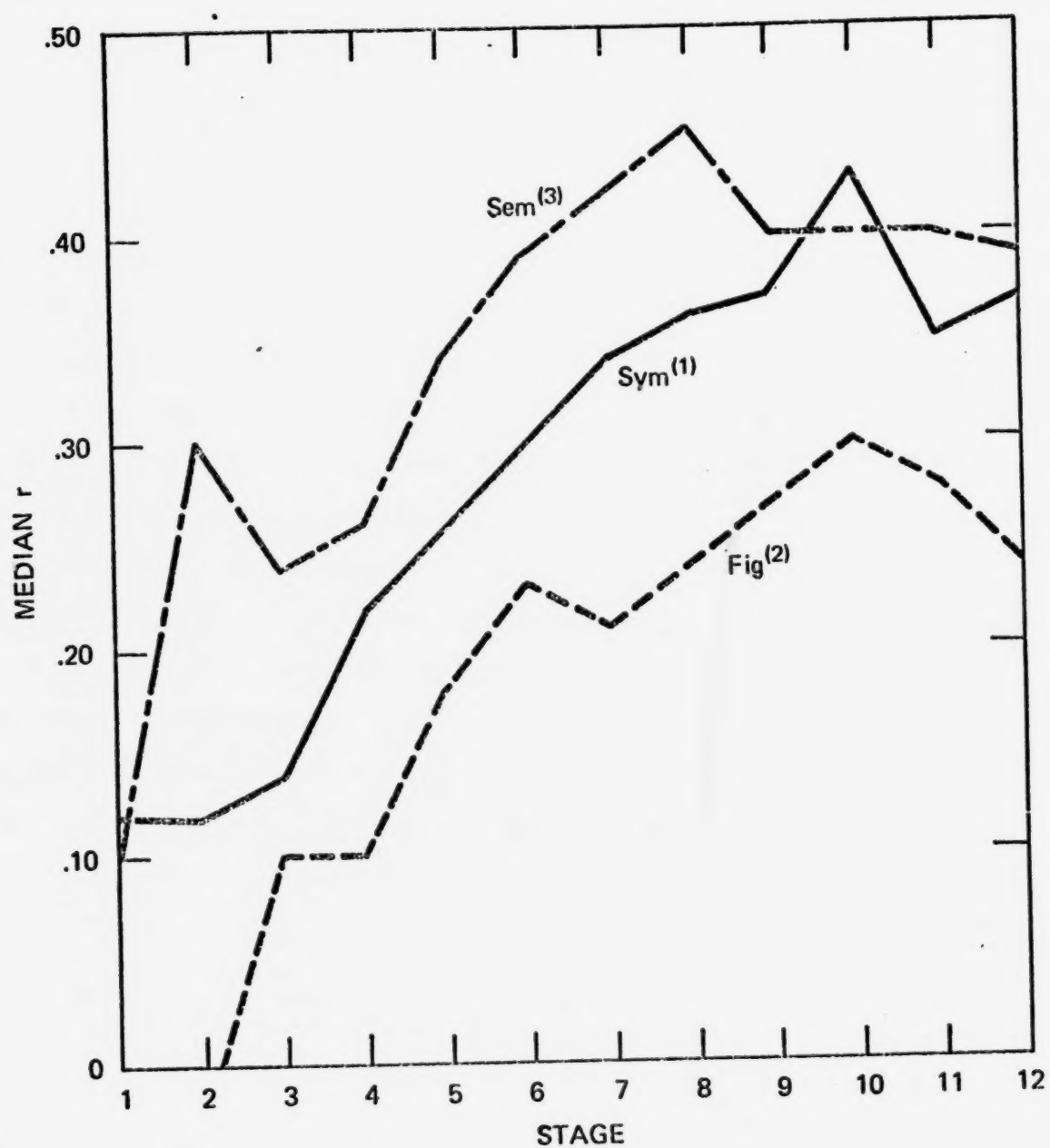


Figure 5. Correlation of concept attainment scores with "general" ability at successive stages (from data of Dunham et al.).

and that an equation fitted to all scores at once (ignoring the earliest problems) would predict any trial about as well as an equation fitted to the specific trial (again, when crossvalidated). That is to say, once initial irregularities are out of the way, learning-to-learn seems to be rather well accounted for by a general ability.

The concept of stages of learning also enters prominently in the several papers of Gagné and his associates on hierarchical structures. One of these is particularly related to individual differences in learning. Gagné and Paradise (1961) outlined a hierarchy of content to be presented through a linear program. It was supposed, as in other writings of Gagné, that "abilities" would be primarily relevant to mastery of the early stages of the program and the relevance to performance in later stages would be less. The chief measure of "learning rate" was the time spent by the student on each segment of the program; this is somewhat ambiguous, as students were free to work superficially if they chose, and thus rapid progress may have meant little persistence rather than quick mastery. Another difficulty is that performance on early trials could be rapid for pupils who had already mastered the subskills introduced there.

Five abilities were measured: vocabulary, speed of symbol discrimination, computational speed, associative memory, and following directions. It was assumed that the first two were irrelevant but it turned out that all tests save vocabulary had very nearly equal correlations with overall scores on learning and transfer, clustering around 0.50. The authors' Figure 3a (p. 12) suggests that computation and associative memory have strong correlations with learning rate on early trials and much lower correlations later (because task-specific learning sets are the chief cause of late-stage individual differences). The data for rote associative memory are impressive and reasonable. The trend for computation, however, is anchored only by two subtests which themselves call directly for computation. There is no reason, then, to interpret this as pertinent to early "learning".

The other part of the stage hypothesis is that correlations of attainment on early sections of the program will be increasingly predictive of performance on later sections. This seems likely, however, to be in part an artifact since all or nearly all pupils mastered some of the early subtasks; consequently, the correlations of those tasks with later learning on the phases next in succession would be low. Later, when there is more differentiation on the achievement measure, it would have greater predictive power. On the whole, then, while the data support the concept of a hierarchy, the hypotheses about individual differences seem not to be confirmed.

Particularly impressive evidence of the pervasiveness of general ability in learning tasks is found in the work of Taylor and Fox (1967), who devised learning tasks at each of the Gagne levels of complexity of learning, from stimulus-response to principles. All tasks were reasonable approximations to military training activities. For five of six tasks there was a marked relation of learning to a general classification test. The exception was a simple monitoring task that required alertness but could scarcely be said to involve learning. As the analysis was a contrast of extreme groups (with a middle-level group sometimes added), it is difficult to put the result in correlational terms. Evidence of separation of the two groups was found at all points in a series of practice or training trials, but ceiling effects typically entered so that the dull came closer to the bright in the end. The statement is made that those who did poorly on one kind of learning did poorly on all, i.e., that over this range of tasks learning ability was general. No statistics on this are given.

Interactions of ability with programming parameters

The first obvious question in an empirical search for interactions is whether general intellectual ability interacts with treatment variables. To organize this report, we need not put a fine point upon the definition of general ability. There is a spectrum of tests ranging from those with substantial educational loadings to those that are intended to be independent of particular training and experience. The more "fluid" tests generally correlate substantially with the "crystallized" tests. We shall lump all kinds of intelligence and scholastic aptitude measures, considering distinctions in the few places where pertinent data are found.

The topic now before us is an ancient one. The recurrent proposal to group pupils by ability rests on the premise that the groups will be treated in different ways, and that consequently their outcomes will be better than if all had had the same instruction. All proposals to isolate slow learners and give them special treatment likewise assume that the special treatment will produce better results for this group. More recently, enthusiasts for programmed instruction have contended that the small-step principle will make the dull student nearly as efficient as the able student. Because this claim stirred up a sizeable amount of relatively well-controlled experimentation in the last decade, it seems reasonable to begin this review with the topic of programmed instruction (PI).

The studies sometimes use a single treatment and report correlations of outcome with a pretest of some kind. Sometimes there is a PI treatment and a conventional instructional treatment for a control group. Sometimes there are two or three contrasting PI treatments (e.g., small step vs. large step).

Overt response as a variable. The studies of PI vs. text, or of overt responding in the programmed mode vs. covert responding (i.e., reading the program with the blanks filled in) are quite diverse, since instructors bring in auxiliary variables and complex designs.

Because of the scattered nature of the studies, we offer the reader the following brief catalog of our conclusions in the ensuing section. The symbol PI refers to the overt mode of practice.

Investigator	Correlation of success with aptitude?	Interaction?	Remarks
Feldman	Inadequate report	Disordinal. PI poor for low verbals	Poor analysis
Wittrock	Not in PI; strong with covert response	Yes. PI good for lows	
McNeil-Kieslar	Yes	No	Overt best
Della-Piana	Yes	No	Possible attitude effect
Scharf	Yes	No	
Lublin	Yes	No	Negative r with autonomy
Williams		No	Poor analysis
Williams	Yes	Negligible	
Burton-Goldbeck	Yes	None overall	Refined, theory-based interaction

A study by Feldman (1965) is so complexly designed that most of the report seems worthless. Three instructional programs in college psychology were formed, at three levels of readability (Flesch). Different pretests and post-tests were given to those studying each set of materials, and raw gain scores were processed. No sensible comparison can be made, since the outcome measures are dissimilar, gain scores treacherous, and ceiling effects likely. While there was a transfer test, uniform for all S's, that apparently gave a fair basis for comparison, a large number of cases with low ability had to be dropped from the analysis and we are given no information as to their distribution over treatments. Comparing persons high on SCAT Verbal with the Low cases not dropped, Feldman finds the High groups nearly equal on overt and covert modes (results from all selections pooled), and the covert mode definitely better for the Low group. Scores were highest with the easy study material, and adding difficulty seemed to impede Lows but not Highs. The regression

lines do not cross. Perhaps the most important finding here is that programmed instruction (overt response required) seemed to hurt the Low-Verbals.

Wittrock (1963) gave a program on science to first and second graders with overt and covert responses. Comparing posttests of those with high and low MA, he found the Lows helped by overt responding and the Highs slightly handicapped by it. The regression of outcome on MA was essentially flat for overt mode and fairly steep for the covert mode; the interaction is significant, though it vanished on a retest one year later.

Another study at the primary level taught word recognition (McNeil & Kieslar, 1963). One group of children observed each frame while listening to the sound track; a matched group of 13 children made an oral response at each point. The mean scores were as follows:

	Low IQ	Medium IQ	High IQ
Overt oral response	25	32	34
No overt response	17	24	30

The oral-response procedure is best at all levels. The interaction is not significant, but the sample size is small.

Della-Piana (1962) developed four versions of a program for college students, the first being a constructed-response version, the second giving a hint as to the response to be constructed (initial letter), the third program giving the response but requiring the student to copy it, and the fourth program using constructed response with feedback and requiring the student to try again if he had prepared the incorrect response. The American Council Psychological Examination (L score) was available as an aptitude measure. Grade averages were also available. The standard deviations of outcome variables are reasonably similar from treatment to treatment, which permits us to interpret correlations directly. There was essentially no difference in the correlations of the four treatments with the L score. There was a lower correlation for grade-point average in treatment 1 than in the other treatments, but with 50 cases or fewer per cell this appears to have been fortuitous. We conclude/that treatment differences did not show any interaction (nor was there a main effect).

A somewhat complicated further analysis employed a semantic differential measure of attitude toward programmed learning. This had a small (significant) correlation with outcome under the constructed-response treatment, a borderline correlation for the fourth treatment (also using constructed response),

and an insignificant relation (on the negative side if anything) for the second and third treatments. That is to say, the persons who began with a favorable attitude tended to respond better to the constructed-response treatment than to the covert-response treatment or the constructed-response with hints. This finding can be given little weight in the absence of confirming experiments. In any event, the magnitude of the interaction is too small to arouse much interest.

Scharf (1961) arranged four kinds of practice by providing varying amounts of feedback in programs on logic. Subjects were required to write an answer to each question, but in the middle sections of the program, correct answers on every item were given for feedback purposes to one group. The other groups had feedback on every other item, feedback on 50 percent of the items in mixed order, or feedback on 25 percent of the items in mixed order. Though there was a strong main effect for IQ, the published data make it clear that there was no interaction of posttest with IQ. In terms of number of errors made during the program, a low rate of reinforcement was clearly disadvantageous to the low IQ group; this did not handicap them particularly on the posttest. For both IQ levels, 100 percent reinforcement produced the best scores.

A study of college students by Lublin (1965) also employed several levels of active responding. Interestingly, it employed the EPPS Autonomy score as an aptitude variable. In different subject groups, 0%, 50% (alternate items), 50% (items selected at random), or 100% of the frames were left to be filled in by the students. The highly redundant Holland-Skinner programmed text was used. Success correlated 0.46 with a scholastic aptitude score (all cases pooled). There was not a significant interaction between amount of responding required and aptitude. (We are not given descriptive information on cell means or regressions.) The autonomy score was correlated (negatively) with attainment -- to the author's surprise, but not ours. Lublin did not find a significant interaction of her several programs with autonomy. This suggests another experiment comparing programmed instruction with a conventional text, with the hypothesis that there will be an interaction with autonomy. Mention may be made, in this connection, of Doty and Doty's (1964) report that sociable students (identified by questionnaire) do worse in PI than others. More evidence is needed to support the implication that it is the nonsociable student, if anyone, who should be assigned to PI.

Another way of demanding greater overtness of response is to present completion as opposed to multiple-choice items. Williams (1965) made this comparison for two groups of college students. While most effects were weak, there was a supposedly significant interaction calculated on raw gain scores divided by working time. This surely is meaningless, judging from the means given for the test scores.

The similar study by Williams and Levy (1964) contrasted constructed-response training with covert response (straight reading of filled-in text). There were four groups, as some subjects had review content. The correlations of verbal ability with posttest were very strong, and with small groups the variation among the correlations has to be regarded as nonsignificant. The pretest should have been used along with the aptitude scores to give meaningful information on the author's correlational questions.

Williams' earlier and more satisfactory study (1963) employed four levels of response (completion, multiple-choice, reading with key response words underlined and straight reading). The latter two were definitely inferior here for the Holland-Skinner material. The two forms of active response gave these results:

	Mean	s.d.	r
Completion	23.5	2.1	0.57
Multiple-choice	23.0	3.3	0.23

(The correlations for the reading treatments were around 0.35.) The slopes of regression on standard aptitude score are 0.76 for choice and 1.20 for completion. The lines seem not to intersect within the range, which would seem to argue for using completion with all subjects, but since completion takes longer, this may not be the best conclusion. In any event, the interaction here is small.

A short program of only 35 frames of isolated facts was used in an experiment on various kinds of presentation (Burton & Goldbeck, 1962). Treatment DA used the multiple-choice mode, in which the correct answer was given along with four highly confusable (difficult) alternative responses. In treatment EA the correct answer is given along with four highly discriminable (easy) alternatives. In group NA there were no alternatives given, but the correct answer was given on the feedback frame. A verbal reasoning test was used as an aptitude measure; due to a somewhat unfortunate allocation of subjects to treatments there were sizeable differences between treatment groups on the pretest. There was no overall interaction of method with aptitude.

A complex breakdown (which had evidently been planned as a part of the experimental design) led to a strange interaction significant at the .001 level. Items on the test were classified according to whether the response was a high-frequency word, likely to be high in the subject's response repertoire when he thinks of a category such as "animal", or low in the repertoire.

The following table of data is given, where each number is a proportion of correct answers averaged over all items and subjects:

Response Availability	Aptitude	Treatment		
		DA	NA	EA
Common	Low	.74	.58	.68
	High	.70	.80	.84
Rare	Low	.29	.42	.56
	High	.57	.54	.63

It is to^{be} seen that the high-aptitude subjects have little or no advantage when the item is difficult (highly confusing alternatives presented) and the response is a common one, or when the response is presented in an easily discriminated form and the required response is a rare one. The high aptitude group has considerable advantage in three of the other cells. This interaction is essentially ordinal, if we disregard the very small reversal in the upper left-hand cell. There was of course a significant main effect for aptitude.

The following extracts from the discussion, although not completely clear out of context, are illuminating:

"We have hypothesized that the learning method for group EA with easy alternatives emphasized response training, while the learning method for group DA with difficult alternatives emphasized discrimination training. The test data indicated . . . that responses not strong enough to be elicited easily . . . benefit more from response training as provided in group EA. The advantage of group EA for these responses was substantially greater for Low verbal aptitude subjects. Apparently the response training provided in group EA was most appropriate for low aptitude students learning responses that were not well established in their response repertory.

"We might expect the discrimination training provided in group DA would be relatively more effective for responses that are well established in the response set. The data lend some support to this expectation in the case of low aptitude students . . . for high aptitude subjects the reverse was true.

"It is worth noting that without analysis of subject and response characteristics, the 'findings' for this study would have been simply that no differences occurred among the three methods of learning."

This study introduces a distressing thought that interactions may reverse themselves depending on the difficulty of the content of the particular association to be learned. In this case it was possible to use the theoretical

difference between association learning and discrimination learning to provide a hypothesis for the analysis; it will probably be more difficult to do this in more complex instructional situations. So far as is known this study has not been replicated, yet in its use of concepts from learning theory to explain an ATI it is one of the most sophisticated studies considered in this summary.

Other studies bear on the same overall question regarding the properties of PI with overt response, contrasted with simple reading of a filled-in text. The claim that PI would yield results uncorrelated with general ability is simply not true. Admittedly, there might be a particular program that reliably gave equally good results for the dull and the bright, but that would be an anomaly in the general run of positive results. This confirms our conclusion that general ability correlates with learning rate, and not merely because the bright person has a headstart on the new material.

As to interactions, the verdict is essentially negative. Three studies suggest presence of an interaction, but sometimes the Lows seem to do best with overt response and sometimes with covert. Any claim of an interaction in this domain must be rejected until it is crossvalidated on new samples and preferably at several ranges of ability. The only finding here that encourages further investigation is the Burton-Goldbeck hypothesis.

"Smooth" vs. "Rough" Programs. Enthusiasts for programmed instruction repeatedly voice the idea that programs that minimize confusion and error enable persons with limited insight to acquire knowledge as fast as others (Stolurow, 1964, 1966). The structure provided by the program, through its use of small steps and orderly sequence, is perhaps not required by the able learners who could organize for themselves, but overcomes a barrier for the dull. Such was the hypothesis. It has been tested over and over, and each study afforded an opportunity for ATI to appear. The experimental variable was sometimes step-size, sometimes smoothness of sequencing. For simplicity we shall refer to both small-step and well-ordered programs as relatively smooth. Attention may be drawn to Briggs' (1968) discussion of the ambiguities of the concept of sequencing in instruction, suggesting that variation in sequencing is not in itself a sufficient basis for defining experimental treatments.

Again, a catalog can be placed before the reader.

Investigator	Correlation of aptitude with success		Interaction	Remarks
	Smooth	Rough		
Maier-Jacobs	Yes	Yes	Between groups; rough superior for abler <u>classes</u>	Not signif. by this calculation. None within groups.
Cartwright	Yes	Yes	Weak; rough superior for abler pupils	Based on our reanalysis. Borderline significance
Smith	Weak	Weak	Weak or none	Residual gains as outcome.
Levin-Baker	Yes	Yes	Rough superior for abler pupils	Not significant; small sample.
Hershberger	Yes	Yes	Possibly	Poor analysis
Campbell	Yes	Yes	No	
Traub	Yes	Yes	No	Unusually competent analysis
Payne <u>et al.</u>	Mixed	Mixed	Some indication that smooth favors abler.	Effect arises from greater internal consistency of learning from smooth sequences.

If further evidence were needed to support the earlier summary, this reaffirms that conventionally-measured abilities are definitely correlated with ability to learn from PI. The studies differ too much, and effects are too weak, for us to draw a firm conclusion about interactions, however. Occasionally, a simplified, smoothly progressing, well-structured program seems to be more effective with duller individuals and groups, but such findings are hardly consistent at this point.

In an excellent large-scale inquiry, Maier and Jacobs (1966) compared PI with instruction involving a teacher in Elementary Spanish. Another large study had the same general style: a year-long series of brief lessons, with two pretest measures (IQ and attitude to Spanish) and three posttests (achievement, attitude to Spanish, and attitude to PI). Seventeen classes had a small-step, orderly program; 22 classes a scrambled program with no regular progression of frames. In this case we shall give a complete table. One reason is that, after studying the reported information which took the class as sampling unit,

we realized that more was to be learned by examining data at the level of the pupil. Dr. Paul Jacobs of ETS kindly calculated additional figures, which appear below the previously-published information in Table 1.

There is no significant interaction. For cases pooled there is no effect. There is a large difference in slopes, strong in the between-groups analysis, but it does not reach significance. Effects of this magnitude may well be important, even though it will always be impractical to employ enough groups to get significance for between-groups effects of this size. The effect arises from the large between-classes s.d. for achievement under the scrambled program. This, one speculates, is a morale effect, some classes being stimulated by the scrambled version and others frustrated. For the ablest classes (judging by averages), the scrambled program appeared positively beneficial. This is not easily reconciled with the slight tendency of these able classes to disfavor PI, whereas with the smooth program it was the high-achieving classes that favored PI.

The means in Section I of the Table are similar. The standard deviations show much larger class-to-class variation in the achievement posttest where Version B was used. (It should be remembered that this is a standard deviation for class means and not for individual students.) Among the correlations (Section II) there are a few differences. Abler classes had a more positive attitude toward programmed instruction than duller classes, when exposed to the small-step version; there was a small effect in the opposite direction for the scrambled version. This seems to be inconsistent with an interpretation offered at a later point in our discussion for results of Stallings and Snow, Salomon, and Koran, where it is suggested that abler students tend to be bored by and to resist methods that require routine attention and give them no latitude for their own reorganization. These studies are not entirely similar to the Maier-Jacobs study, however. It is interesting to note that the classes where ability and achievement were highest were also those where attitude to PI was most favorable, under the small-step treatment -- but there was no particular relation between class attitude and class achievement under the scrambled treatment.

TABLE 1

Results of the Maier-Jacobs Experiment on Instructional Programs

I. Means and standard deviations of class means		Version A (Small step, orderly)		Version B (Scrambled)	
Pretest variable		<u>Mean</u>	<u>s.d.</u>	<u>Mean</u>	<u>s.d.</u>
Kuhmann-Anderson IQ		106.5	6.6	103.5	8.3
Spanish attitude		4.5	.5	4.7	.7
Posttest variable					
Achievement		20.5	4.7	20.2	7.8
Spanish attitude		10.5	1.1	10.3	1.6
Attitude to PI		.5	.2	.5	.3
II. Correlations and regression slopes calculated between classes					
	Correlations for each version*			Regression slopes for each version*	
	<u>IQ</u>	<u>Sp (pre)</u>	<u>PI</u>	<u>IQ</u>	<u>Sp (pre)</u>
IQ		-.01 -.02	.75 -.32		
Spanish attitude (pre)	-.01 -.02		-.08 -.03		
PI attitude (post)	.75 -.32	-.08 -.03		.02 -.12	-.03 -.13
Spanish attitude (post)	.04 -.19	.50 .71	-.11 -.01	.01 -.04	1.1 1.6
Achievement	.70 .82	.07 .21	.74 -.27	.50 .77	.7 2.3
*A above in each pair, B below. In regression, variables are treated singly					
III. Results calculated for pupils singly					
	<u>Mean</u>	<u>s.d.</u>	<u>Mean</u>	<u>s.d.</u>	
IQ	107.1	12.1	104.3	13.3	
Achievement	20.8	11.0	20.9	12.9	
Correlation		.59		.64	
Slope, achievement on IQ		.53		.62	

The original report gave sufficient information for calculating the regression slopes for class means. The most striking result here is that the regression of achievement on tested ability is considerably steeper in the scrambled treatment. This suggests a significant and practically important interaction, but the Maier-Jacobs analysis, following strict statistical logic, took classes as the sampling unit since classes, not individual students, had been assigned randomly to treatments. The results were treated as if there were 17 cases in one group and 22 in the other. In order, then, to find out whether these apparently large effects would have been reported as significant by a more conventional analysis using students as the basis for analysis, we asked Dr. Jacobs for a further report. He supplied the information in Section III of the table.

Again we see that the mean difference is small. Standard deviations for individual students turn out to be the same under the two treatments. This, taken together with large differences in standard deviations in Section I, suggests that treatment B generated marked between-class differences, upsetting the treatment-A correspondence of mean ability to mean attitude. The correlations of ability with achievement are reasonably comparable to those given in Section II, except that, as is usual, correlations calculated on the individuals are lower than those calculated on group means. The regression slopes for all cases pooled differ negligibly. A desirable further analysis would be to examine the within-class information, since the data in Section III combine within- and between-class effects.

Our final conclusion is that the difference between treatments produced no clearly significant ATI. On the other hand, the between-class effects are large, and one difference in between-class slopes was impressive. That effect could not reasonably be expected to reach significance, even in this experiment of exceptionally large size, but the finding is large enough to be of practical importance. The implication seems to be that one runs extra risk of poor learning in applying the scrambled version to a class which on the average has low ability, and risk of low morale in applying the scrambled version to a class of high ability!

While we do not understand this study, we find in it a significant hint that interaction effects may be mediated by the reaction of a class as a whole, and not necessarily at an individual level. That is to say, a class with certain characteristics may be more prone to profit from one treatment than another, and this opens a way to recognize pupil variation just as important as the individual assignments that ATI studies usually have in view.

Cartwright (1962) taught mentally retarded adolescent students fractions by means of two programs, one arranged to show the successive fractions in a systematic, natural manner (e.g. $1/2$ and $1/3$, $1/3$ and $1/4$, $1/4$ and $1/5$, etc.), while the other presented the sequence in an irregular order. There were no differences in immediate learning, but there were differences on retention and transfer tasks. Paradoxically the smooth treatment was superior for producing retention, but the scrambled treatment was superior for producing transfer. Cartwright also reported correlations between his criteria and several pretest variables including general ability, language, prior knowledge of arithmetic fundamentals, and simple mathematical reasoning. He gave little interpretation and did not look for interactions, though Stolurow (1964) uses the study's results as an example of interaction. From the correlational report, it ~~had~~ appeared that general ability was correlated significantly with performance on the rough program and not on the more structured one. On the other hand, certain subscores from the achievement pretest seemed substantially more correlated with learning under the smooth sequence. We re-analyzed raw data provided in his appendix to obtain proper tests of that interactional hypotheses. Results are presented in Table 2 for our re-analysis of the raw immediate, retention and transfer criteria. Cartwright's gain measures have been ignored. F-tests on pairs of slopes yielded only one effect approaching significance: the scrambled program produced a stronger relation between full scale IQ and the immediate posttest.

Smith (1962) did another dissertation study Stolurow cites as evidence for ATI in programmed instruction. The study is a substantial one, with 133 fifth graders in various groups. There were three kinds of programmed instruction, with more or less fine steps and more or less strong prompts. In addition, there were cases receiving regular classroom instruction. Seven concepts about fractions were taught. The aptitude measures included PMA and Guilford divergent-thinking tests.

The dependent variables were handled in a quite unsatisfactory way. A "criterion" test was given three times, and simple residual gains were calculated as learning measures. Correlations were then calculated between the residuals and the aptitude tests. (The Tucker-Damrin-Messick paper criticizing this kind of correlation had not been published in 1962.) The second major

TABLE 2
Results of Reanalysis of Cartwright Data

<u>Aptitude Measure</u>	<u>Criterion Measure</u>					
	<u>Immediate Posttest</u>		<u>Retention Test</u>		<u>Transfer Test</u>	
	r	slope	r	slope	r	slope
Pretest	.72	.58	.66	.42	.55	.34
	.49	.50	.57	.53	.56	.30
Arith, Reasoning	.68	5.12	.48	2.86	.51	2.89
	.64	4.88	.60	4.17	.48	1.94
Arith. Fund.	.74	7.14	.67	5.12	.54	3.99
	.41	3.23	.44	3.16	.23	0.98
Total Arith.	.73	6.52	.59	4.16	.55	3.71
	.55	4.38	.54	3.92	.37	1.58
Total Reading	.47	3.47	.62	3.62	.63	3.54
	.57	5.40	.64	5.63	.43	2.21
Total Language	.40	2.58	.51	2.58	.52	2.53
	.63	6.67	.40	3.90	.24	1.37
Total Grade Placement	.56	4.30	.60	3.67	.60	3.47
	.62	6.25	.57	5.30	.38	2.03
Full Scale IQ	.19	.28	.33	.39	.38	.43
	.61	1.57*	.16	.38	.27	.37

For each set, figures for smooth program appear above, for scrambled program below. *indicates slope comparison has $p < .10$. $N=20$ and 16 for smooth and scrambled programs, respectively.

fault is that correlations rather than regression slopes are interpreted. There are some rather large differences in s.d. within groups on both pre-tests and posttests, and examination of slopes is called for. The information reported on correlations is essentially useless, since only "significant" correlations are given. But there were fifteen ability scores, correlated with several outcome measures in eight groups, and any single group had 20 or fewer cases; this means that the scattered "significant" correlations very likely reached that level by chance, and that consistent but nonsignificant correlations were suppressed. The essence of the Smith conclusion is that programmed and conventional instruction produced different correlations of aptitude with outcome; moreover, the pattern of correlations varied at each stage of learning and with different kinds of outcome. We are unable to accept or deny this on the basis of the statistical analysis presented. This is a particularly distressing example of a painstaking and laborious experiment rendered worthless by faulty analysis.

Time did not permit us to reanalyze the study, although Smith did report all her raw data. One tiny analysis will serve to reinforce our comment on the undependability of Smith's interpretations and Stolurow's several secondary accounts of the study. Says Smith (p. 120) in one of the few entirely straightforward generalizations of the report: "The total ability test score was related only to performance under a conventional method of instruction". And on the same page "total general mental ability" is tabled as "involved in" conventional and not programmed instruction. If we look back for the source of this generalization we find that two outcomes (both residualized) were correlated with Total PMA score in each of eight treatment groups. The first four were classes taught by different live teachers; the first outcome score has no significant correlation; one out of four correlations were significant for the second outcome. None of the eight correlations for sections having programmed instruction was significant. From this contrast Smith generalizes.

We recalculated, pooling all classes having the conventional instruction, and also pooling two classes both of which had the third of her programmed sequences. These are the key results (Program A being the "rougher" program):

	"Definitive"		"Inferential"	
	Question posttest		Question posttest	
	r	slope	r	slope
Conventional (N=77)	0.13	0.011	0.29	0.010
Program A (N=15)	.10	.005	.22	.024
Program B (N=13)	-0.17	-.034	.17	.013
Program C (N=28)	.02	.003	.18	.017

(The reader must not dismiss small slopes out of hand; they partly reflect the scoring scales chosen.) Even though the correlations for the conventional instruction are higher, the difference is not at all significant. The slopes, more important for interaction, give a hint of steeper slope in the conventional treatment for the first outcome and the opposite for the second. (Yet it was a correlation for the second outcome that initially led Smith to her generalization.) We recommend that the Smith conclusions be disregarded unless a full multivariate reanalysis is made.

A 17-day series of PI lessons for second graders by Levin and Baker (1963) contrasted an organized and/^ascrambled order of steps. There were only 18 cases and the report is too incomplete for interactions to be properly examined (e.g., no s.d.'s reported). There is a hint of disordinal interaction such that the scrambled program is best for those with high IQ, but the effect is not significant. Here, results under both the smooth and rough program correlated substantially with IQ.

A different kind of "smoothness" entered the Hershberger (1964) study; the original program was redundant and an edited, terse version with all redundancy stripped out formed a second treatment of the material. Both redundant and terse programs were presented in two versions: with and without quiz questions. A California Reading Test score was used as aptitude measure. Results were definitely better with quiz questions. Performance was strongly related to aptitude in two treatment cells, moderately in the other two. There is some indication of interaction: Given the discursive text, adding questions was far more helpful to the High aptitude group; given the format without quiz questions, the terse text was better for the Highs than the discursive text. Results are not well reported; raw gains are used instead of a proper regression on the pretest and aptitude measures, and there seem to be some

inconsistencies between the text statements and the published figure.

A study that produced no interaction worthy of attention did demonstrate^a clear and consistent positive correlation of learning from programmed instruction with aptitude (Campbell, 1963). Since set theory was taught, a mathematical ability test was used as pretest. There were short, long, and "bypass" programs, the latter including branching directions to shorten working time for the student who mastered a subskill. On the whole the longer (smaller-step) programs taught somewhat more. Neither high, medium, or low aptitude groups exhibited consistently different trends over the three program formats.

The possibility that homogeneity or heterogeneity of practice material would differentially influence student learning was entertained in a study (Traub, 1966) involving three days of instruction on the graphical addition of positive and negative integers on the number line. The first two days of instruction were common to all groups (approximately 100 sixth graders in each group), but on the third day the problem sets were heterogeneous for one group, homogeneous for the second group, and for a control group a mere time-filling activity. A total of 36 aptitude measures were collected, either by giving factor tests or using achievement tests from the school record. There was a main effect favoring the heterogeneous problems group. Of primary interest here is the fact that Traub tested the homogeneity of regression slopes and accepted the null hypothesis; that is to say, there is no significant interaction. Although there were 36 aptitude variables, a factor analysis reduced them to twelve, so as to make it less likely that power would be lost through the sacrifice of degrees of freedom. This was distinctly desirable, though a full multivariate test (rather than tests of slopes for several predictor variables singly) would have been still more powerful. Despite this criticism, technically this is one of the most excellent studies in the literature on ATI.

A common section of the instructional program was correlated with the various aptitude test scores. Both rate and error scores correlated substantially (ca. 0.50) with verbal and reading measures, and slightly less with measures in arithmetic. Within the differentiated programs, correlations were basically similar for all treatments. There were large correlations for an arithmetic pretest and modest correlations for a composite of reasoning and quantitative tests. Why the tests that best predicted the common parts of the program failed here remains mysterious.

The Payne, Krathwohl, and Gordon (1967) study tried eight versions of a program teaching statistical concepts to college students. It was thought that the "more scrambled" versions would show more relation to ability, but the correlations (with an arithmetic test) varied erratically. If the authors had examined regression slopes they would have found evidence contradictory to the hypothesis; though the trend is weak. The completely scrambled version had a slope about one-seventh that of the most scrambled. The reason is that the correlation of scores on part-tests were higher for smooth sections of the program than for rough sections. Smooth programs produce more score variability!

Miscellaneous PI studies. A contrast of pure PI with a combination of programmed and live instruction has the excellent features of the Maier-Jacobs study already discussed.

Maier and Jacobs (1964; see also Jacobs, Maier, and Stolurow, 1966, p. 60 ff.) examined the effects of a year-long televised series of PI lessons in Spanish. Some classes had PI only, some had PI plus lessons from a teacher, and a third group had live lessons without PI. There were three aptitude variables: Kuhlmann-Anderson IQ, Spanish pretest, and Spanish attitude pretest. There was a significant main effect for treatments (PI - only being distinctly inferior) and a significant main effect for ability. The interactions found significant were these: favorable attitude toward Spanish (an outcome) was associated with teacher-plus-PI instruction for High-IQ classes, and with PI-only or teacher-only instruction in Low-IQ classes. Second, there was an ATI with attitude toward PI as an outcome; this was interpreted as showing that low-ability students tended to favor PI and reject live teaching; while high-ability students tended to favor live teaching. Also noteworthy was the fact that attitudes toward Spanish were highly stable in the teacher-plus-PI group between pretest and posttest, and that under the other treatments the pretest-posttest r was negligible. Taking attitude as an important outcome then, teacher-plus-PI is good for students with favorable pretest attitudes, not for those with unfavorable initial attitudes. The further fact that some teachers got better results than others produced higher-level interactions. All this led to the recommendation that High IQ students should have teacher-plus-PI instruction from a teacher who favors innovative methods, and low IQ students should be taught by a live conventional teacher, without PI. Information on cell means and standard deviations is too scanty to indicate how much

practical value there is in this assignment policy.

The many excellent features of this study should be noted: reproducible treatments (at least insofar as the PI component was concerned); use of two treatment dimensions (extent and character of live teaching); use of more than one aptitude variable and use of more than one outcome variable; instruction extending over a realistically long time with real class material; and a large sample (77 classes with some 900 students). It will be noted that here again the investigators employed the class as the sampling unit, whereas nearly all educational investigators have employed persons as the unit. This is technically correct, but it means that the authors have labelled nonsignificant many effects that would have been called significant by most of those who report on these matters. Their nonsignificant effects are often large, and if confirmed with more classes could be important. No doubt a cumbersome multivariate analysis that examined both between-class and within-class regression would extract more information from a study with so complex a design; but the state of our insight into the interaction is not far enough developed to make good use of the statistics toward which the field will surely move.

A small study on a related matter compares experimental subjects taught the Arabic writing system by programmed drills with others trained by a live teacher (Carroll and Leonard, 1963). The results clearly favored the experimental method. There was a strong correlation of achievement with the Modern Language Aptitude Test (0.70 within treatments, pooled). But the regression slope was virtually identical for the programmed treatment and the live-teacher treatment.

An interaction showing steeper slope for PI than for conventional instruction was reported for instruction in English usage (Reed & Hayman, 1962).
(1964)

A study of some importance by McNeil/unfortunately stands without replication. It deals not with ability as a differential variable but with sex. (We have not tried to pull together the research on sex differences, though the reader can find much in Maccoby (1966) that is tangentially related to the ATI problem. None of it relates directly to instruction.) In the McNeil study, there was autoinstructional training in reading (individually administered by machine) in kindergarten. Then there was first-grade teaching by a female teacher using conventional methods; seven teachers were involved with different subjects. Whereas boys and girls had been equal on a pretest, the boys were a bit superior after the PI phase, and dis-

tinctly inferior after four months in first grade. Both differences are significant. (Correlations with IQ and pretest were not discussed.) The author speculates that PI reduces the distracting social activities of boys, and he gives some evidence that the live teacher does not give boys as much opportunity to learn as girls. While the hypothesis that automated instruction serves boys best is noteworthy, the confounding of variables in the design of this study make the finding equivocal.

Silberman et al. (1962) undertook to develop distinct programs that would appeal differentially to "overachieving" and "underachieving" students. A fairly lengthy pair of programs in geometry was prepared and pretested. One pilot study sought an effect of program difficulty, presence of anxiety-arousing statements, and test anxiety, in combination, on achievement. The only detectable difference was that the easy program under anxiety conditions elicited better performance from the Low anxious and poorer performance from the High anxious. With six cases per cell this could not be taken seriously.

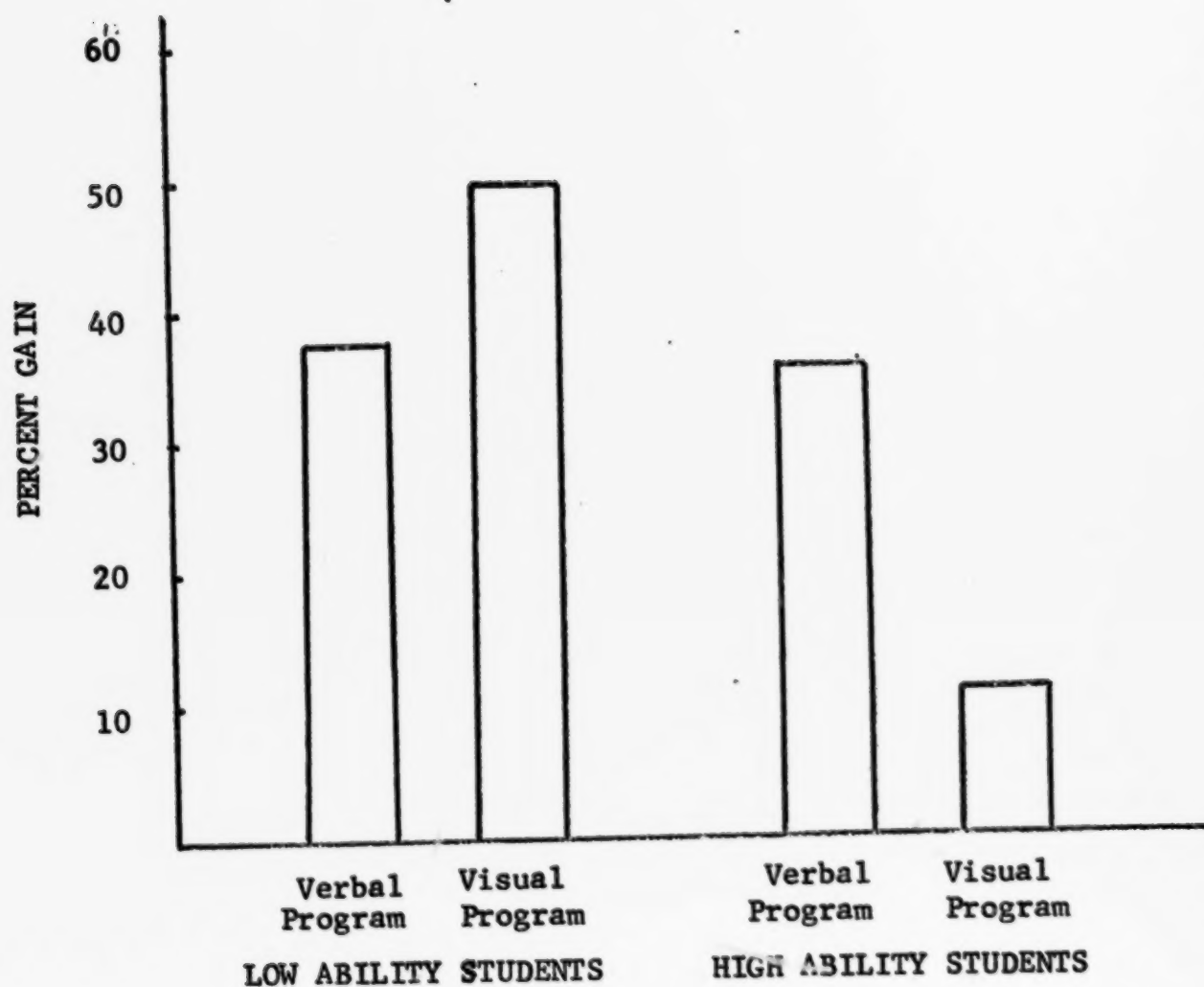
The main study presented proofs of seven theorems, one in a rote, step-by-step manner, one in a "conceptual" manner with considerable discussion of method of attack. The two programs produced very similar results. Over-achievers, normal achievers, and underachievers differed in precisely the same manner under both programs. The sample was sizeable. A more powerful analysis would have treated mental ability, base achievement measure, and pretest on the program as covariates. But no reanalysis could have produced evidence of interaction in these data. It had been expected that over and underachievers would do equally well on the more interesting "conceptual" treatment, and would differ under rote instruction. But there were differences in both treatments.

Two findings of interaction involving general ability are reported by Taylor and Fox (1967). While the studies did not use formally programmed instruction they are close in conception to some of those above. In one, a number of military symbols used in map-making were taught. Method A allowed the trainee to study symbol meanings from cards, by his own devices. Method B was a controlled practice in which the stimulus was presented, the subject responded, and the correct response was given. The order and timing were controlled. Method A worked better for all men, but was particularly superior for abler men. In a complex-plotting task the methods contrasted were: Method I, a television presentation with small-steps including pictorial examples, and practice on specimen problems with knowledge of results; Method II, lecture, practice without knowledge of results. Basically, Method I simplified the task of the learner. Method I was indeed superior, and strikingly so for the

dull learner. The interaction is ordinal, implying that Method I is superior for all men. But it costs more, and Method II is highly efficient for abler men. There is a strong ceiling effect in the results, but on this task it seems unlikely that extending the task to more difficult problems would make Method II scores superior.

A pair of studies by Gropper (1965) seemingly led to a practically important interaction. But a close reexamination questions the finding and raises major questions about current standards of reporting research on instruction. Figure 6 is a crude reproduction of a chart from the AIR Annual Report for 1965; we regret that we cannot display in all its impressiveness the report's artistic rendition in color. The reader who has followed our methodological discussions will have serious reservations about the reporting of raw gains, but he will not realize the following further facts: Two posttests were used, and only one of them showed an interaction. As nearly as can be judged from the report's analysis of variance for gains, there was no hint of interaction for the second (pictorial) test. In a second study of much the same nature, there was no interaction for the immediate achievement test. There was an interaction, roughly like that from the first study, of an achievement test given four weeks later. But this was true for only half the subjects (one of two orders of presentation); for the other half of the group, the interaction effect was apparently negligible. Too few descriptive data are presented to permit a solid evaluation of the results. One sees a progressive selection in this report. Analyses of variance are presented only when there is some significant difference in the tables, and many of these are relegated to an appendix. The text does not always indicate that the data in its tables are for a restricted subsample. And the abstracts of the studies and the dramatic Annual Report give no hint at all that the interaction effect showed only in highly selected analyses.

The study itself is neither more nor less solid than most of the studies we have reviewed. Junior-high-school students were taught principles of physics either by visual displays on closed-circuit television or by programmed verbal materials (also presented on TV.) On a verbal posttest the means for pupils with low and high IQ (test not identified) were 13.2 and 15.7 for the visual presentation, implying a modest relation to ability; and 12.8 and 17.8 for



Gains in achievement for low- and high-ability students who learned Archimedes' Principle by means of demonstrations emphasizing verbal and visual presentation.

Figure 6. Demonstration of reporting from American Institutes of Research Annual Report, 1965

the verbal presentation, implying a strong relation. The interaction would have been heightened if a pretest were used as a second aptitude. The interpretation of this result is made bewildering by the fact that there were eight, not two, treatment groups, involving such combinations as repetition of lesson with active response, shift from verbal to visual lesson with no active response, and single lesson with active response. It is impossible to figure out which subgroups entered into the comparison of means that led to an interaction. Presumably the active response variable did not contribute to an interaction or that would have been mentioned. We do know that there was no interaction for the verbal posttest.

As for the second study, it was if anything more complicated, and the analysis less clearly reported. On the delayed achievement test only, in half the cases only, IQ was steeply related to performance in a treatment that called for verbal responses and very little related in a treatment where the pupil responded by marking one of three pictures in his booklet.

If some of these results were to be replicated and brought under control so that they would appear regularly, they might indeed be important. The probability that the verbal-pictorial difference will interact with general ability is thrown into some question, however, by the internally inconsistent findings of the Gagne-Gropper (1965) study to be discussed later, though the treatments in that study were of a different character.

We have not attempted a detailed review of the possibility of interactions where television and conventional instruction are contrasted, although an overview and discussion of some of this literature was provided by Snow and Salomon, 1968; Technical Report No. 3. On the basis of our experience in other areas, we would warn the reader against accepting any author's conclusion without careful reconsideration of the statistical treatment. As Campeau (1967, Pp. 106-107) summarizes the findings in this area,

Differences in effectiveness between instruction by TV and by conventional methods have sometimes been found to vary with ability level. High-ability students learned significantly more by TV than by conventional methods in psychology (Dreher & Beatty, 1958) and in science (Jacobs & Bollenbacher, 1959); . . . conventional methods of instruction were most effective for low-ability students in science (Curry, 1959, 1960; Jacobs & Bollenbacher, 1959) However, conventional instruction was significantly better than TV for high-ability learners in English composition (Buckler, 1958) and mathematics (Curry,

1959) while TV was significantly better than conventional instruction for low-ability learners in economics and psychology (Dreyer & Beatty, 1958) and mathematics solving (Jacobs, Bollenbacher, & Keiffer, 1961)."

Whatever the final verdict on the original studies, the correct conclusion is evidently that the difference between television and conventional instruction is not a variable about which one can generalize.

One study of televised instruction deserves separate mention since it deals explicitly with differences in pacing conditions (a treatment variable left implicit in most TV vs. live comparisons as well as many other studies of alternative instructional methods). Kress and Gropper (1966) produced 12 separate versions of a televised program to form 3 levels of prompting and 4 levels of tempo or presentation rate. Subjects were matched on IQ and independent estimates of their characteristic work rates in self-paced instruction. The study was not designed to test ATI, but a reorganization of the data led Kress and Gropper to an important ATI hypothesis. They observed that (p. 277),

"Subjects who were characteristically fast under self-paced conditions out-performed the characteristically slow Ss only when both worked under fast, fixed tempo conditions. The superiority of the fast workers was evident in all six criterion measures. Surprisingly, however, in five of six measures, the opposite effect was observed under slow, fixed-tempo conditions. Here, the fast workers committed more errors and achieved lower scores than did the slow workers. Thus, though fast workers were matched with slow workers for IQ, under the slow fixed-tempo, where impairment would not be expected to be great for either group, fast workers did more poorly.

"The general pattern, then, . . . , revealed that mean performance was highest when characteristically fast students worked under a fast fixed-tempo, and when characteristically slow students worked under a slow fixed-tempo. Lowest means resulted when characteristic work rates and externally controlled tempos were not matched."

It is unfortunate that these investigators could not pursue this hypothesis more directly in their data. Regression analyses using both IQ and characteristic work rate to predict criterion performance could have shown more clearly the nature of a pacing interaction and also reduced uncertainty about the effects of imperfect matching. In view of its relation to hypotheses reviewed elsewhere in this report, this study clearly requires replication in an expanded form.

A study conducted by Woodruff, et al. (1965) on pacing conditions in programmed instruction serves as another example of research on aptitude-treatment interaction using totally insufficient statistical analysis. Eighth-grade students (N=74), receiving TMI-Grolier's complete programmed course, General Science, were randomly divided among teacher-regulated and self-regulated pacing conditions and again among in-class and out-of-class use conditions. The resulting four treatment combinations are described in Table 3. At the end of each of two semesters, posttest performances and the number of correct frames were recorded as criterion data. The program consisted of 3,053 frames for the first semester and 3,469 for the second. Data collected and analyzed in the original study include scores on a total of 40 variables: 21 scores from Torrance Tests (fluency and originality measures); grade average in all school subjects; grade average in science; Gates reading speed, vocabulary, comprehension and total reading scores; Lorge-Thorndike verbal and nonverbal raw and IQ scores; pretest on science; 1st posttest criterion score (end of first semester); 2nd posttest criterion score (end of year); 1st semester gain (posttest 1 minus pretest); 2nd semester gain (posttest 2 minus posttest 1); year gain (posttest 2 minus pretest); number of correct frames, separately for each semester and for total year.

There were no significant differences in subject-matter achievement or number of correct frames among the four treatment groups. Woodruff found only scattered significant correlations of the 21 Torrance measures with achievement. There were significant correlations of previous grades, intelligence, and reasoning with most measures of achievement. The relation of mental ability and first-semester achievement, however, generally did not hold for second-semester achievement, perhaps because of a decline in achievement among high-ability students. A less favorable attitude toward PI during the second semester was also reported. Nothing in this analysis bears directly on the ATI Woodruff intended to study. Using raw data provided in an appendix of the earlier report, E. I. Sawin and Snow reanalysed, computing a series of simple regression analyses on predictor-criterion pairs separately for each of the original treatments, to show the extent to which disordinal interactions are apparent among regression slopes. A complete intercorrelation matrix including all variables was computed for each of the four treatment groups. Such matrices had not been provided in the original report.

TABLE 3

Woodruff Study Treatment Groups

	IN CLASS	OUT OF CLASS
OWN RATE	<u>Group I</u> (N=15): Programmed material used during regular class period. Student progressed at own rate of speed. Individual tests and conferences initiated by student when he completed a unit. Expected to meet a minimum time schedule. Supplementary work assigned if student finished program before end of semester. Also homework, but not programmed	<u>Group III</u> (N=16): Programmed material assigned as out-of-class work. Students progressed at own rates, but a minimum time schedule was established as in Group I. Unit exams and conferences with teacher during regular class periods initiated individually by students. Supplementary work assigned for class time not needed for exams and conferences. Students who finished program before end of semester were assigned further supplementary work.
TEACHER REGULATED RATE	<u>Group II</u> (N=25): Programmed material used during regular class period. Rate of progress regulated by teacher to a great degree by having regularly scheduled times for discussions and unit exams. Individuals who completed a unit ahead of schedule were given supplementary work. Supplementary materials also assigned as homework.	<u>Group IV</u> (N=18): Programmed material assigned as out-of-class work. Students progressed at own rates, but regularly scheduled discussions and examinations were initiated by teacher as in Group II. Class time not needed for discussions and tests was used on supplementary materials and activities.

The correlations for all possible pairs of independent and dependent variables were examined for all four treatment groups to identify those pairs in which at least one correlation differed from zero using $p < .05$. Of 256 pairs (32 aptitude variables crossed with 8 criteria), 103 showed significant correlation in at least one of the four groups. For each of these pairs, a comparison of regression slopes among the four treatment groups then was conducted. Of the 103 pairs, 20 were found in which the overall F -test of slope heterogeneity exceeded the .05 level. This number represents approximately 8% of the original 256 pairs. Scatter plots and regression equations were obtained by computer for each of these 20 pairs of aptitude-criterion variables. The bivariate distributions appeared to justify the linearity assumption as well as could be expected in groups of 15 to 25. Multivariate analyses were not attempted, though results would clearly be more simply interpreted by combining predictors.

Among the 20 regression analyses, there are several interesting clusters of findings, though almost all significant ATIs involved the number-of-correct-frames criterion. The results may be summarized as follows: For Torrance's Consequences (both originality and total scores) and Improvements (fluency, originality, and total scores) tests, high positive relations were obtained in the self-paced, out-of-class condition. In other treatment groups, these slopes were negligible. Thus, high scorers correctly completed more of the program under out-of-class treatment while low-scoring individuals performed best in the self-paced, in-class condition. There was some tendency for those lowest on Improvements and Tin Cans flexibility to do best when placed in the teacher-paced, out-of-class treatment. Considering the intelligence and achievement measures, high positive relations were obtained between criterion performance and Lorge-Thorndike scores in all treatments except the self-paced, in-class condition, where the corresponding relation dropped to zero (see example in Figure 7). In practice, these findings would suggest that students with IQ scores below 111 be assigned to self-paced, in-class conditions while students with scores at or above 111 be assigned to self-paced, out-of-class use. Using grade-point average as an aptitude measure, results similar to those reported for Lorge-Thorndike scores obtain for individuals in the lower half of the GPA distribution. Those above that point would be assigned to the teacher-regulated, out-of-class condition if

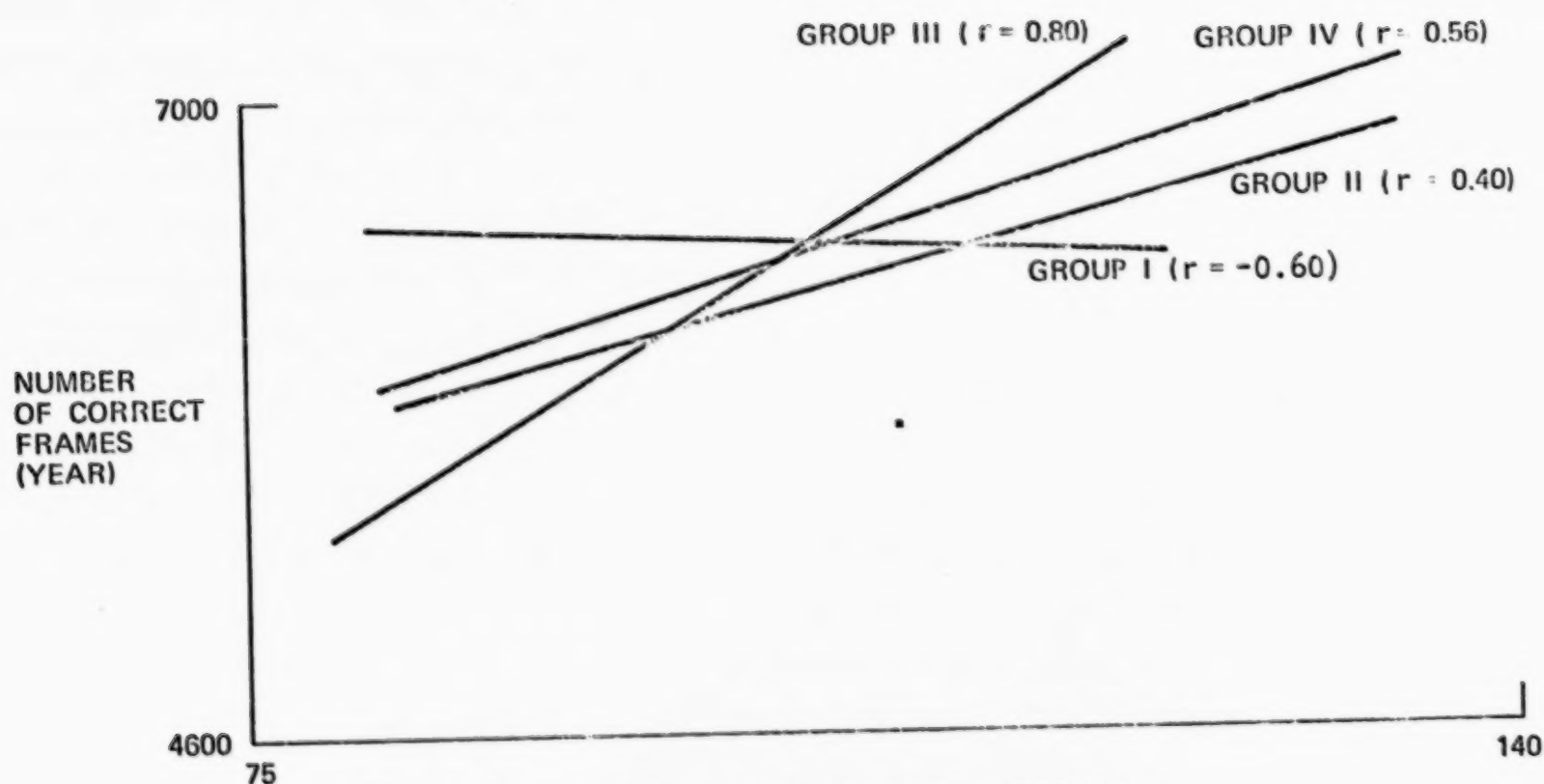


Figure 7. LORGE - THORNDIKE, VERBAL IQ

total grade-point average is used to predict a criterion of correct program frames, but to the teacher-regulated, in-class condition if grade average in science is used as the predictor of posttest criterion performance. A similar finding was obtained using Gates reading speed as the aptitude dimension.

The only result involving a criterion other than number of correct frames was obtained using prior grade-point average in science. The criterion was raw gain in the second semester. Highest criterion scores were obtained by those with high science GPA if they were assigned to teacher regulated-in class work. For low GPA students self-regulated work in class seemed best.

Meaningfulness of instruction as a source of interactions.

One of the oldest, and one of the newest, hypotheses about ATI is that rote and meaningful instruction will serve different kinds of students. In the older literature, this emerged as an incidental rather than a deliberate finding. In the most recent, learning theory becomes the basis for postulating two radically different kinds of learning and two associated kinds of instruction.

This view is most explicit in the numerous writings of Arthur Jensen (see, for example, 1969) in which he puts forward the concept of distinct Level I and Level II abilities. The Level II abilities are thought of as much like the fluid ability or *g* of other writers; they involve deliberate analysis, employment of meaningful interpretations (mediation), and self-criticism and correction. The Level I abilities are employed in tasks that call for rote associational learning -- essentially, tasks such as digit span and serial learning. These are defined negatively, as tasks that do not lend themselves to mediation. (The concept of crystallized abilities enters Jensen's system obliquely; for him, tasks at either level may have greater or less educational loading.) A paired-associates task might be a Level I or a Level II task, depending on whether the person brings mediation to bear. Jensen has not organized systematic data to support his conceptual separation of Level I and Level II, though he does refer to a number of studies that are at least tangentially relevant. The proposal is reminiscent of the repeated separation of "memory" factors in test research; among the studies we have reviewed, Allison's gave the clearest evidence for a rote memory factor. That research is not entirely suitable for checking on Jensen's proposal, since subjects in test research have never been tuned to use mediation where it is appropriate; consequently, a given sample is likely to use a mixture of styles and to give unclear results. What is a rote task for one subject is analytic for another. But the problem is capable of direct investigation.

Jensen, having made the separation, takes a step unthinkable in educational psychology since about 1930 -- he proposes that instruction be deliberately carried out on a rote basis for some children. More drastically, he argues that good Level I ability is characteristic of Negro children and that this is probably hereditary, so that he is proposing a radical form of educational segregation. Assignment to treatment would be made indi-

vidually, of course, but it is clear that he expects ghetto children to predominate in the rote-taught classes. This is not the place to argue the philosophy of such a proposal or the pertinence of the evidence on heritability (but see Cronbach, 1969). Here the need is for a survey of evidence on ATI with meaningfulness as an instructional parameter.

In the 1930's, in a reaction against Thorndikean emphasis on drill and under the correspondingly welcome influence of Gestalt psychology, many studies were done to demonstrate the superiority of "meaningful" over "rote" instruction. Information on individual differences in response was not of particular interest, and only scattered reports are available. The most noteworthy is that of G. L. Anderson (1941), a doctoral dissertation conducted under T. R. McConnell at Minnesota and published in a short form. Pupils were assigned at random to conventional instruction emphasizing practice, or to instruction that made a deliberate attempt to develop meanings. Instruction continued for a year, and a large sample was tested, making this an exceptional study for those times or ours. Pretest data were an arithmetic test and a test of general ability. A number of outcome measures were taken, and the data were processed by the rare but appropriate Neyman-Johnson technique. Clear "regions in significance" emerged where the meaningful instruction gave best results -- consistently over all outcomes -- for one type of student and the less meaningful instruction best for another type. As can be seen from the figure prepared by Cronbach (1967) from the unpublished dissertation, the data would lead one to assign to the less meaningful instruction those who have hitherto been "overachievers" in arithmetic, and to assign to the meaningful instruction those who show good general ability and poor performance in arithmetic. We do not have at hand information as to the steepness of regression slopes that would indicate the power of such a decision rule to improve instruction.

This finding can be given a commonsense interpretation. Instruction up to this point had been relatively meaningless, in this school in the mid-1930's. Where a pupil has done well in past work, he presumably has some study techniques or more basic aptitudes that make him a good prospect for further instruction of that sort. Where a pupil has done worse in arithmetic than an ability test implies he should, an alternative treatment sounds like a good investment. We cannot argue that the suitability of meaningful

instruction is directly a matter of "Level II" ability. We suspect that Anderson's mental test was a hodgepodge with considerable v:ed loading. Nor can we argue that the improvement came from the cognitive suitability of the new approach for these children; perhaps it was simply more interesting for pupils who, on the evidence, had not been responding well to the old approach. What is especially to be noted here is that a separation was made within the broad aptitude domain we have hitherto been referring to as "general". In the research on PI summarized above, the pretests on which interactions were to be based might be haphazardly chosen from the crystallized or the fluid sector; we know of no study where a deliberate attempt to separate out past achievement from fluid ability was made. In a few studies (e.g. Smith) there is multivariate pretest information that would permit a reexamination along these lines, but such a contrast seems never to have been planned. Nor, when mental tests are used, has serious thought been given to contrasting v:ed and fluid subtests. We note also that Anderson's results bear only tangentially on the hypothesis that Level I abilities have relevance to ATI.

A few years after the Anderson study Brownell and Moser (1949) reported their magnificent study, carried out in dozens of schools, of meaningful vs. mechanical instruction in subtraction. A brief account of the main themes of the study is given by Cronbach (1963, pp. 342-344). What concerns us here is an incidental finding. In half the schools, subtraction was rationalized for the children; a major effort was made to explain why certain steps were performed in (e.g.) borrowing. But third graders in some of the schools seemed unable to profit from these explanations. The authors tell us that where instruction had been rote in the two preceding grades the whole concept of explanation in arithmetic was strange to these pupils, and they could not incorporate the meanings offered. These children, then, had developed a positive inaptitude for meaningful instruction, whereas other children had been led to the point where they could profit from explanation. Now this is important first in undermining the concept that aptitude or readiness is simply a matter of intellectual maturity. Second, it sharply challenges such a concept as Jensen's regarding a native incapacity. Third, it destroys any lingering attempt to define "one best way" of instruction. Fourth, it urges us in the direction of trying to help the pupil who does not use

meaningful instruction effectively by combining techniques that will move his skills forward without relying on comprehension, with techniques that will advance his ability to comprehend. We are in no position to write off these third graders as noncomprehenders -- but we do not anticipate that simple tuning will bring them to the level of mathematical reasoning.

Brownell and Moser had not planned on a study of individual differences and offer only anecdotal evidence on this overwhelmingly important ATI. So far as we know there has been no follow-up on their research, though the theme of developing comprehension has been present in all later discussions of elementary instruction in mathematics. Indeed, there has been a remarkable absence of research on what constitutes meaningful instruction in various subjects, and what makes a pupil able to profit from it. The work of the experimental and developmental psychologists on mediation has been remote from the educational problem, and has used such short-term treatments as to be of little use. Here, then, is a major field where ATI studies can profitably combine with an attempt to understand quite basic processes of intellectual development and performance.

We may mention incidentally a line of work opened up as a dissertation problem by one member of the project staff. (Project funds paid for some of the computer costs, even though the study is not immediately in the ATI terrain,

because of the ultimate interest the theoretical advance in this direction would have and also because reading performance was directly under examination in the project work discussed below.)

Pearl Roossinck Paulson set out to write a computer program that would "simulate" the comprehension of textbook paragraphs by intermediate-grade children. The child is asked to summarize the paragraph; the computer is set the same task. If the method succeeds (by a Turing test) the human and computer outputs will be indistinguishable. (Actually the present output of the computer is in a telegraphic language that has to be edited before the comparison is made). A program has been written, data have been collected, and a study of the adequacy of simulation is in progress. Basically, the technique suggests that comprehension at this level requires a stored hierarchy of concepts so that specific information can be fitted into the structure; it is obvious also that this recoding process is itself a significant skill or aptitude. But individual differences were not the focus

of the Paulson study; research on performance at different developmental levels would be a reasonable sequel.

To return to the general theme of meaningfulness, we may recall that some of the work on PI could be seen as pertinent. In the category of "smooth" vs. "rough" comparisons, several offered a sequenced treatment to some pupils and a scrambled sequence to others. Surely the authors thought the scrambled sequence was less meaningful. In other studies, "small-step" and large-step programs were compared; here, the smaller steps are in one sense more meaningful. Insofar as there were hints of interactions, it appeared that the rougher, less meaningful presentations were more effective with abler pupils. (This will be seen also in the Edgerton study below.) From the vague theoretical conceptions that exist in the PI field one might have expected the abler pupils to overcome much of the disadvantage introduced by large steps and scrambling, but one would scarcely have expected them to learn more from less-structured instruction than from highly organized instruction. The puzzling results can be rationalized in part by noting the possibility that the rough programs were meaningful to a reasonable degree, or else that the material was so disconnected that the claim to meaning in the smooth version was illusory. Indeed, unless there was some sense in the scrambled version one/^{cannot} understand how duller pupils could learn from them at all. We need also to entertain the possibility that the less smooth program is advantageous because it keeps the capable pupil attentive and refreshed. In this argument, both sets of instructional material are potentially meaningful, but the pupil is given responsibility for reorganizing and consolidating ideas in the rougher version. As we have seen, the PI studies were not entirely consistent and only weak effects can be claimed. Our next section deals with related possibilities in this area.

The studies of D. P. Ausubel advance the argument that one can improve a subject's comprehension of lesson-like material by supplying relevant advance organizers, i.e., preliminary texts that provide a frame of reference or apperceptive mass. In the Fitzgerald-Ausubel experiment (1963) two groups were formed and one received a relevant organizer on Civil War material while the other half worked on irrelevant content. Both then studied a Civil War passage. There was no interaction of treatment with a pretest. The groups with organizers were superior on the immediate and delayed test, and those having superior pretest ability did better; but there was no dependable departure from parallel regressions. Analysis

of variance is a relatively weak procedure for this study. Comparing multiple regressions for the two treatments using both pretests and verbal ability would have led to a more certain conclusion about the absence (or presence) of interaction.

A complex program in science was presented under six conditions by Merrill & Stolurow (1966). There were advance summaries and two kinds of review, in various combinations. While there were main effects, neither verbal nor quantitative aptitude produced a significant interaction. No descriptive data related to individual differences were reported.

Interactions of ability with complex method variables

A program of studies initiated by Stallings and Snow within this project, not yet brought to completion, compares alternative instructional methods in initial reading. In a pilot investigation, the aptitude measures were psycholinguistic and memory abilities, represented by experimentally-produced auditory and visual sequencing tests and in selected scales from the Illinois Test of Psycholinguistic Abilities (ITPA). After administration of the aptitude battery, 20 first-graders were divided to form two comparable groups. For the first two months of school, one group received a phonics treatment (PH), using Arnold Fires materials, while the other group received a look-say (whole word) treatment (LS) using Scott-Foresman materials. The look-say method is, in conventional terms, more meaningful. Two teachers (unaware of S's aptitude scores) alternated daily to balance teacher effects. All Ss were observed periodically by observers who made notes of behavior suggestive of learning avoidance. Indicators noted included excessive fidgeting, distracting neighbors, fighting, fooling, chair-rocking, etc. The accumulated frequency of avoidant acts during the reading instruction served as one criterion for the study. This frequency correlated 0.82 with similar observations made by teachers in the children's other classes. At the end of two months, the California Achievement Test in Reading and the Murphy-Durrell Reading Readiness Analysis were administered to provide further criterion information.

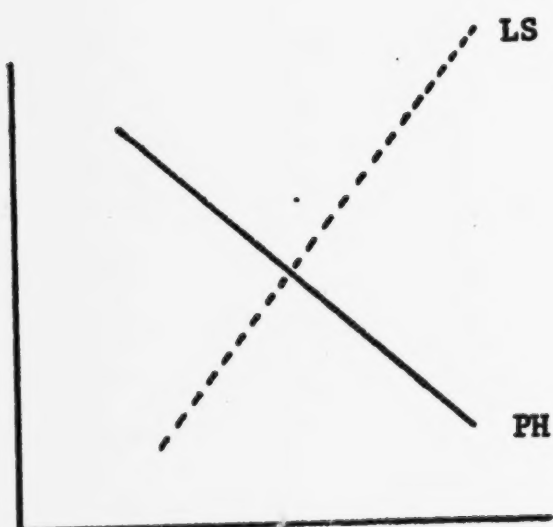
Figure 8 presents simple regression analyses for selected aptitude and criterion measures. The number of cases does not warrant multivariate analysis, nor can significance levels be taken very seriously. Figure 8 shows frequency of learning-avoidant behavior to be differently related to visual sequencing skill in the two treatments.

$$\begin{aligned} LS &= 1.09X - 42.10 \\ PH &= -3.93X + 23.69 \\ F &= 17.02 \text{ (df = 1, 16; p .01)} \end{aligned}$$

Few
Instances

Learning
avoidant
behavior

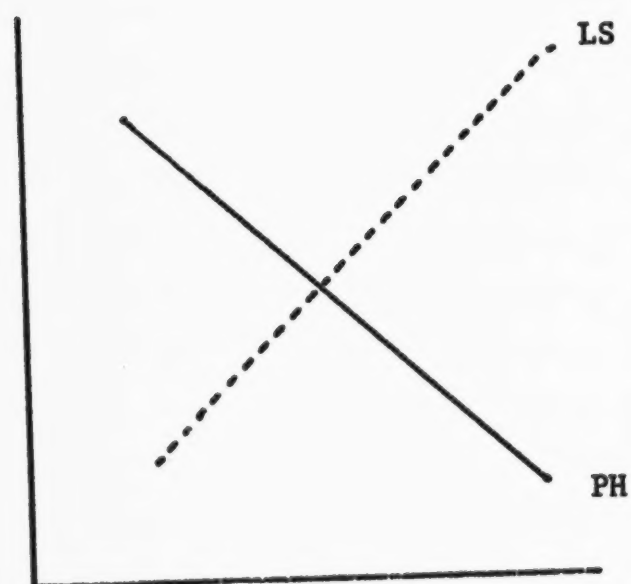
Many
Instances



a) ITPA #9
Visual-motor Sequencing

$$\begin{aligned} LS &= 1.17X + 29.69 \\ PH &= -.61X + 72.51 \\ F &= 5.51 \text{ (df = 1, 16; p .05)} \end{aligned}$$

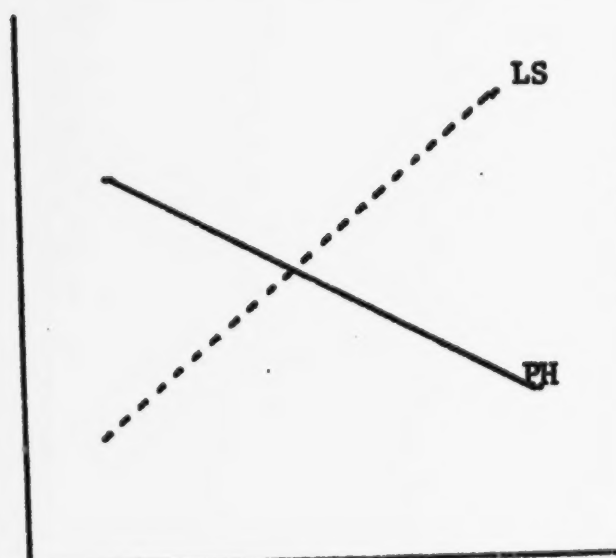
CAT
reading
scores



b) ITPA #8
Auditory Sequencing

$$\begin{aligned} LS &= .91X + 8.18 \\ PH &= -.42X + 11.74 \\ F &= 6.48 \text{ (df = 1, 16; p .05)} \end{aligned}$$

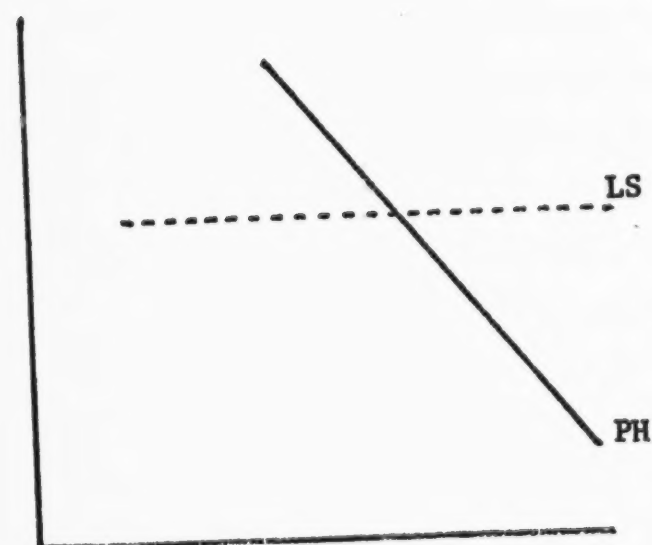
CAT
Reading
Scores



c) Experimental Auditory
Sequencing Test

$$\begin{aligned} LS &= .00X + 17.09 \\ PH &= -1.43X + 57.63 \\ F &= 7.59 \text{ (df = 1, 16; p .05)} \end{aligned}$$

Murphy-
Durrell
Posttest
Scores



d) ITPA #1
Auditory Decoding

Figure 8. Stallings and Snow Study. Selected interactions between reading aptitudes and reading criteria for phonic (PH) and look-say (LS) treatments.

Phonics seems to serve poor visual sequencers best, while more able sequencers seem better off with the look-say method. The remaining graphs, using other dependent and independent variables, give a similar impression. The LS treatment appears to require S to depend on his own sequencing ability and short-term memory. The PH treatment (involving analytical, structured drill) perhaps provides an external substitute for such ability. Among able Ss, however, PH gets poor results; boredom is a possible explanation. Other interactions not reported here also show negative slopes for phonics, but zero slopes for look-say, as in Figure 8J).

These are fragmentary results, based on only 20 subjects. A replication of the experiment has been conducted, but data analysis is incomplete. Here again, the experimental control was terminated after two months so that children not showing progress could be reassigned to alternative or combined groups. If funds can be obtained, a third replication and followup on all subjects is planned.

From other sources also, there is anecdotal evidence that these ITPA subscales relate negatively to reading achievement under phonics instruction. And Bond (1935) long ago suggested that memory span and success in reading under look-say treatment are positively related but negative related under phonic treatment. If these results can be substantiated, it is futile to ask which of these reading methods is the "one best way".

Salomon (1968) divided 26 teachers in training randomly between treatments designed to improve hypothesis generation (HG) or cue attendance (CA) skill. Both groups saw a film made by random reorganization of scenes from a coherent film. HG Ss were then asked to provide as many alternative explanations or hypotheses about the underlying story as they could. The film was replayed until S produced at least 12 hypotheses. CA Ss were asked to report as many stimulus details as possible. Trials were repeated until S had listed 150 visual details. Unlike other studies reviewed earlier, here the stimulus materials were standard but the HG treatment stressed meaning.

After training, a transfer test ("Information search") posed a complex problem involving the development and staffing of secondary-school English departments in a Spanish-speaking, poor district. Each S generated questions he would want to have resolved as he worked on the problem, his score being the number of questions listed. The aptitude variable of interest was general verbal reasoning (GRE-V).

HG training led to more question asking where GRE-V scores were 550 or better (Figure 9); Ss scoring below 550 produced more questions after CA training. CA apparently requires one to lift restrictions on attention, -- report details without evaluating. Perhaps this bores the more verbal subjects. It may be best for less able Ss precisely because it promotes attention to detail. HG training may require more verbal analytical and reasoning skill and thus may be more challenging to high-ability Ss. In other terms, HG and CA approaches may represent alternative problem-solving strategies, each useful to different pupils. Such findings might also result if skill in CA is prerequisite to HG performance, and both are dependent on general verbal facility.

Koran (1969) designed alternative treatments involving various kinds of models to improve intern teachers' ability to ask analytic questions in a microteaching situation. Microteaching is a laboratory arrangement in which the trainee teaches a prepared lesson to a few pupils for ten minutes. The trainee receives feedback or criticism and replans his lesson, then returns for a further teaching trial with other pupils. Teacher interns (N=121) were randomly divided to form three groups. In a video-modeling treatment (VM), S viewed videotape of a master teacher performing the required skill between microteaching trials; here the greatest amount of information is provided to the trainee. In the verbal-modeling treatment (WM), S studied a typed transcript of the sound track for each trial. In the control treatment (NM), S received no information between trials. The number and nature of analytic questions asked by S during each trial, and printed tests of ability to identify analytic questions, served as criteria. Pretest aptitude measures were perceptual and verbal factors from the ETS Kit (French, Ekstrom, and Price, 1962) and Seibert, Reid, and Snow, 1967.

The relation of criterion performance to Hidden Figures scores is traced in Figure 10. The WM treatment worked best for those scoring high on Hidden Figures, while VM was best for low scorers. Hidden Figures performance can be interpreted as an index of general ability, of Thurstone's Flexibility of Closure factor, or of Witkin's field independence. The verbal, self-paced, unrestrictive, articulate treatment, WM, made this aptitude of positive value. In VM, (audiovisual, fixed-pace, attention-restricting), "low aptitude" Ss did best.

$F = 8.24$ ($df = 1, 22$; $p < .01$)

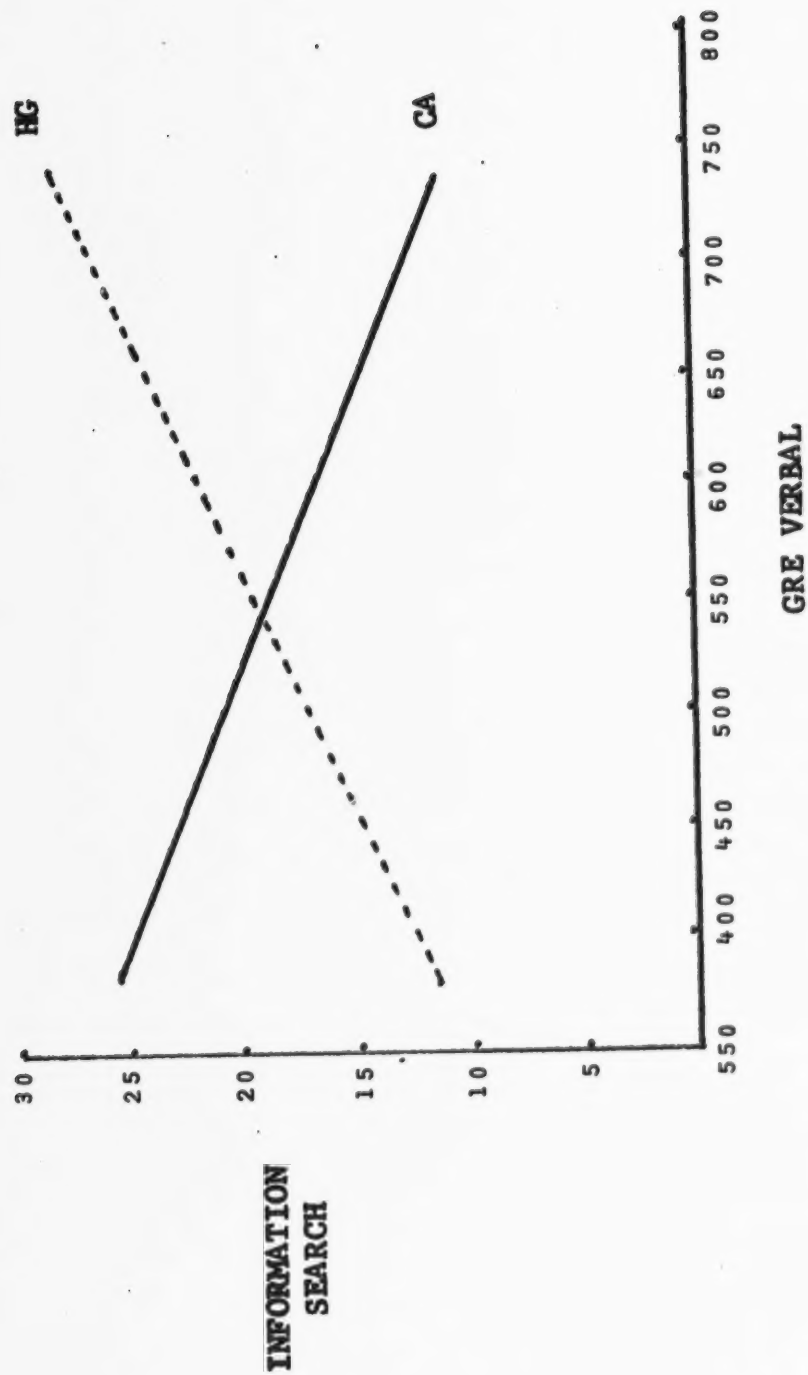
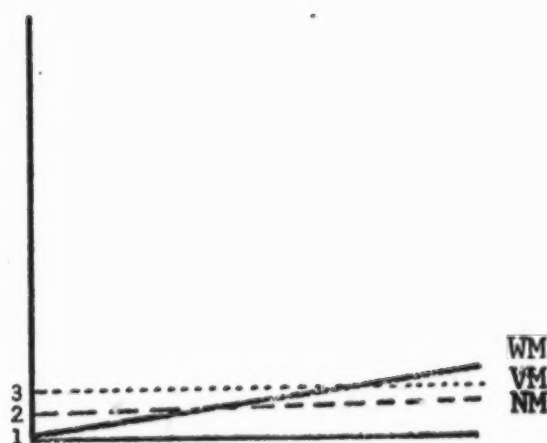


Figure 9. Salomon Study. Interaction of GRE verbal scores with Hypothesis Generation (HG) vs. Cue Attendance (CA) training using information search scores as criterion.

TOTAL
ANALYTIC
QUESTIONS
T₁
(Before
treatment)



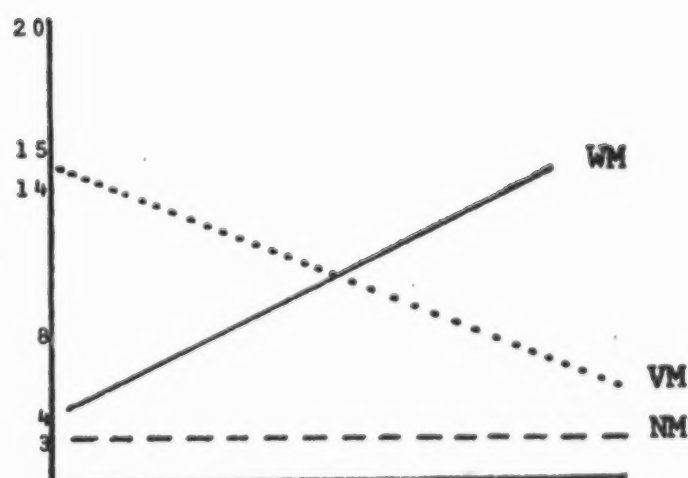
a) HIDDEN FIGURES Part 1

$$VM = .03X + 3.44$$

$$WM = .24X + 2.05$$

$$NM = .14X + 2.90$$

TOTAL
ANALYTIC
QUESTIONS
T₂



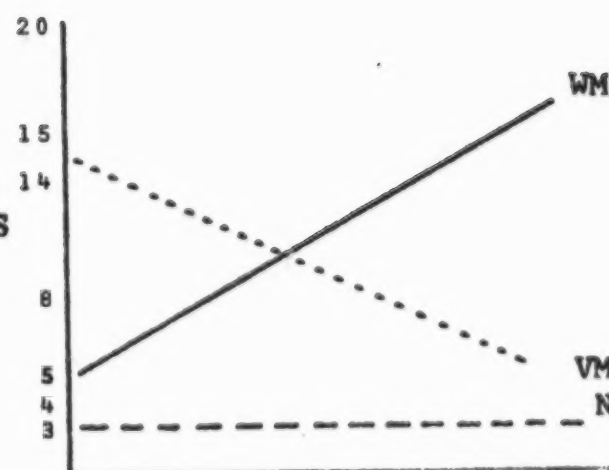
b) HIDDEN FIGURES Part 1

$$VM = -.24X + 14.90$$

$$WM = .61X + 4.19 \quad p < .05$$

$$NM = .02X + 3.34$$

TOTAL
ANALYTIC
QUESTIONS
T₃



c) HIDDEN FIGURES Part 1

$$VM = -.39X + 16.08$$

$$WM = .96X + 3.87 \quad p < .05$$

$$NM = .05X + 2.71$$

Figure 10. Koran Study. Part a), b), c), show separate regressions of performance criterion (total no. of analytic questions) on aptitude (Hidden Figures Test, Part I) for the three teaching trials, respectively. Treatments are written verbal modeling (WM), videomodeling (VM), and control (NM).

A total of 69 regression tests resulted in 13 significant ATI. In addition to the Hidden Figures results, some similar interactions were obtained using Maze Tracing Speed, another perceptual test. Also of note were interactions obtained with an experimental film memory test, to be described later. No ATI were obtained using tests representing perceptual speed, verbal comprehension, auditory memory, or expressional fluency.

One should base hypotheses about individual differences in learning on theoretical frameworks. This was Melton's (1967) view when he distinguished three components in associative learning: stimulus differentiation (where stimuli are identified and coded for internal representation); association (where stimuli-as-coded are linked with appropriate responses); and response integration, where response elements are combined or sequenced in action.

In Figure 11, these components have been collapsed to two. The emphasis of PH, CA, and VM treatments appears to be on attention to and differentiation of stimulus detail. (Any complex instructional treatment would involve all components. In LS, HG, and WM treatments, these processes presumably are necessary, but are not explicitly forced.) Now suppose that the aptitude variables used in these three studies are all taken as aspects of conventional mental ability. Perhaps lower-ability Ss are weak primarily in attentional and discrimination skills, as suggested by Zeaman and House (1967). This would account for weakness on PH, CA, and VM. Perhaps such Ss are deficient also in employing certain kinds of coding during conventional instructional tasks. PH, CA, and VM treatments compensate for this lack by detailed drill that isolates stimulus elements and employs concrete representations. These treatments may provide coding systems for the learner. This perhaps explains the good performance of low-ability Ss. High-ability Ss do badly on these treatments, perhaps because they reflect emphasis on detail. Pertinent to this suggestion is the finding of Wicklegren and Cohen (1962) in an experiment where "memory capacity" was manipulated by allowing some subjects to store information physically; the Ss with the larger memory did poorly because they tried to retain too much detail. Where highs do better -- LS, HG, and WM -- Ss can mediate, abstract, and reason at their own pace. The emphasis in these treatments is on the rapid manipulation of symbolic meaning, probably a preferred mode of operation for high-ability Ss but one unsuited to low-ability Ss. These latter treatments are more similar to conventional meaningful instruction, where outcomes typically relate positively to mental ability.

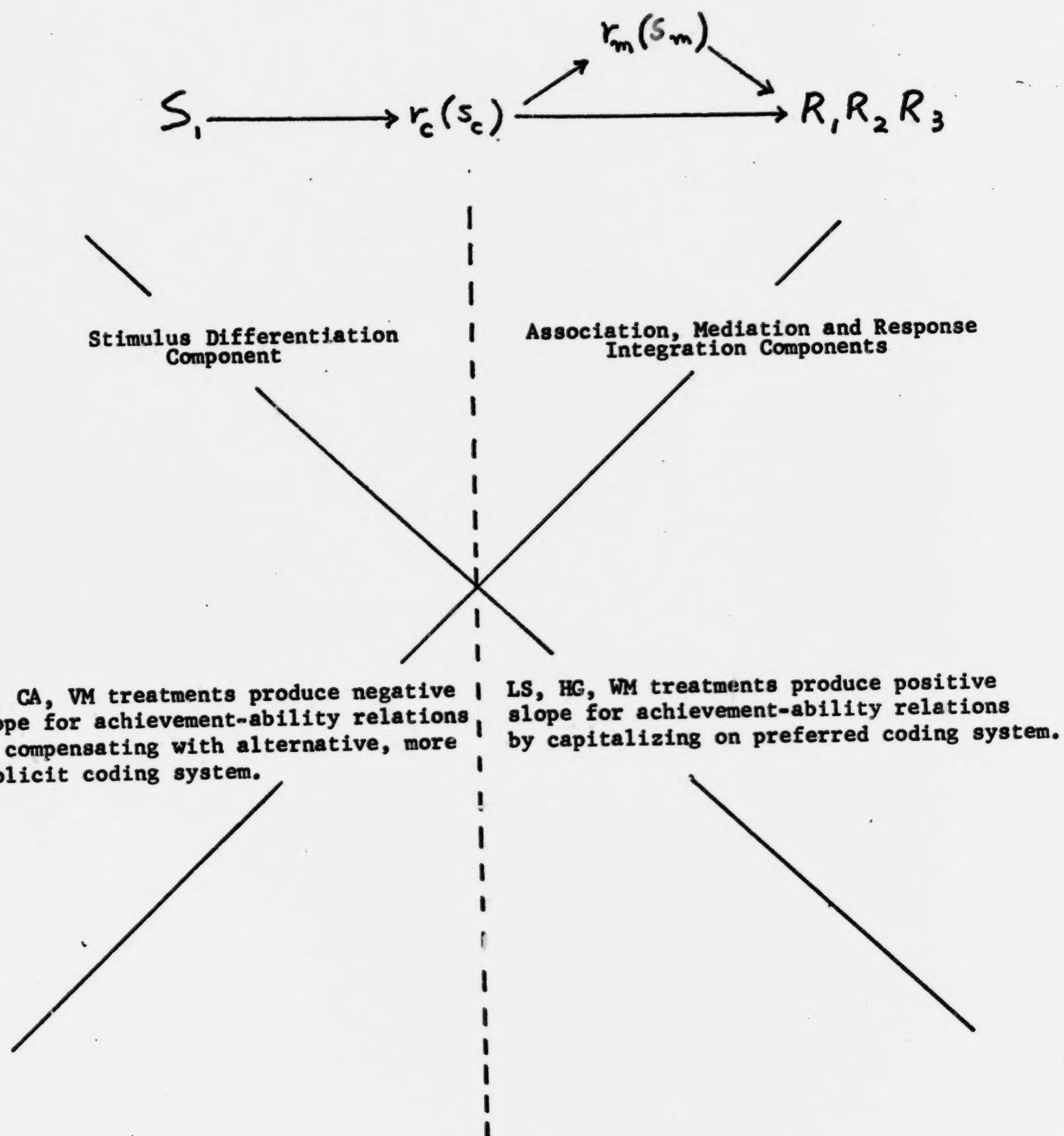


Figure 11. Results of three studies projected onto Melton's model for associative learning.

Thus, it seems that interaction arises in these studies from a compensatory-conciliatory process in which some harmony between the defects or style of the learner and the characteristics of the treatment is achieved. For the low-ability S, the good treatment provides aid in stimulus differentiation, where he is deficient. For the high-ability S, the good treatment provides associational latitude.

If this hypothesis has validity, it should be possible to take some further steps toward expanding the current conception of aptitude. Those who do best under PH, CA, and VM treatments should be describable in positive terms, not merely as persons of low ability. We should be able to construct new aptitude measures, which tap skill in stimulus differentiation and in the use of coding systems not typically found in instruction. Some attempts at such development have resulted in a number of rather crude experimental tests using motion pictures as the communication medium (Seibert & Snow, 1965; Seibert, Reid, & Snow, 1967). One such measure, called Film Memory, was involved in the Koran study. In this test, S attempts to recall the content of a live-action silent film. The content portrays complex human behavior, largely nonverbal, that is probably hard to encode in symbolic or verbal terms. Combining this test with Hidden Figures in regression analysis, we get Figure 12. The criterion here is the number of analytic questions asked by teacher trainees on trial T₃ of the three-trial micro-teaching treatment. (The control group is not shown.)

The planes for the two modeling treatments are differently pitched. Film Memory functions in a manner opposite to that of ability measures considered earlier; the regression slope is positive for VM and negative for WM. The two planes intersect in a line, which has been projected onto the base plane to show how a two-dimensional decision rule would be used to classify Ss on the basis of aptitude. Ss high on Hidden Figures and low on Film Memory should be given the written modeling treatment. Those with good Film Memory, especially if poor on Hidden Figures, would be better off in the videomodeling treatment. The earlier discussion applies here. One can interpret positive slope as indicating that a treatment capitalizes on an ability, while a negative slope suggests that a treatment compensates for low ability and frustrates high ability.

Figure 12 suggests that multiple combinations of specially constructed aptitude measures may be necessary to reach large ATIs that generate considerable payoff. The kinds of aptitude defined in the base plane of Figure 12 represent a significant departure from traditional conceptions. Film Memory

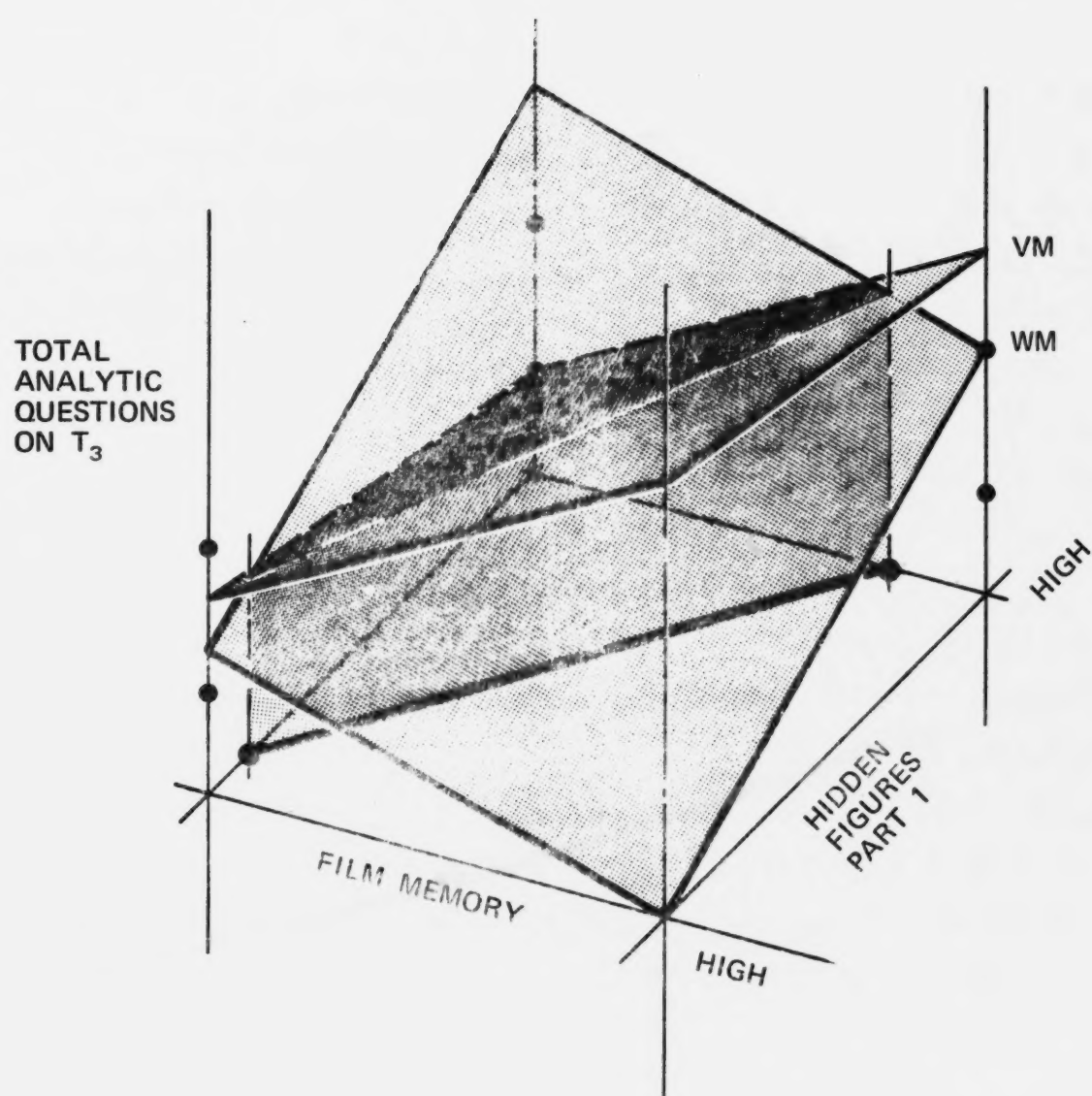


Figure 12. Multiple-regression analysis for Koran study, including video-modeling (VM) and written verbal modeling (WM) treatments.

is not an ordinary experience in test taking, even for college students, and Hidden Figures Part 1 is just the first half, perhaps the learning-trial half, of a conventional test. Possibly these measures represent information processing or learning strategies, styles, states, or sets, and are inadequately interpreted using the static terminology of factors and traits.

The preceding discussion leads toward a general conception of instructional treatments as prosthetic devices for particular aptitude groups. A treatment that proves especially appropriate for a person deficient in some particular aptitude may be functioning as an "artificial" aptitude. It contains the information precessing functions that the learner cannot provide for himself. Whether there is value in this conception as a guide for identifying ATIs remains to be seen.

A study of strategy in verbal learning.

An intricate and technically advanced study by Carl Frederiksen (1967), working with Ledyard Tucker, is worthy of notice even though it is unlike other studies in this summary and cannot be presented in its full complexity. Three groups of college students learned sixty miscellaneous words by an anticipation method. One group followed the serial-anticipation method, one was asked to anticipate the words in groups of five (in their original, haphazard order), and one (free recall) was asked to anticipate the whole list of sixty words. There were eighteen trials. A large number of aptitude scores from the French Kit were reduced^{to} seven composites. In addition, a questionnaire was used to find out what strategy S was using at each point in time. Outcomes and strategies both were processed as a function of time.

The most useful source for our purposes is the author's tables 29-31 which give simple correlations of aptitude composites with learning on successive trials. Figure 13 shows, first, the correlations of the Associative Memory score with learning under each condition on each trial, rather crudely smoothed. The correlations are high and significant, and evidently differ with the treatment and the stage. For Semantic Spontaneous Flexibility (divergent) tests, the correlations have remarkable differences not only in magnitude but in sign. With 40 cases per group, this might be fortuitous; but there is a strong hint that divergent thinking and clustering are antithetical. Verbal ability and memory span have steady correlations with

outcome around 0.35 under free recall, and negligible correlations under the other treatments. Finally, associational fluency correlates significantly

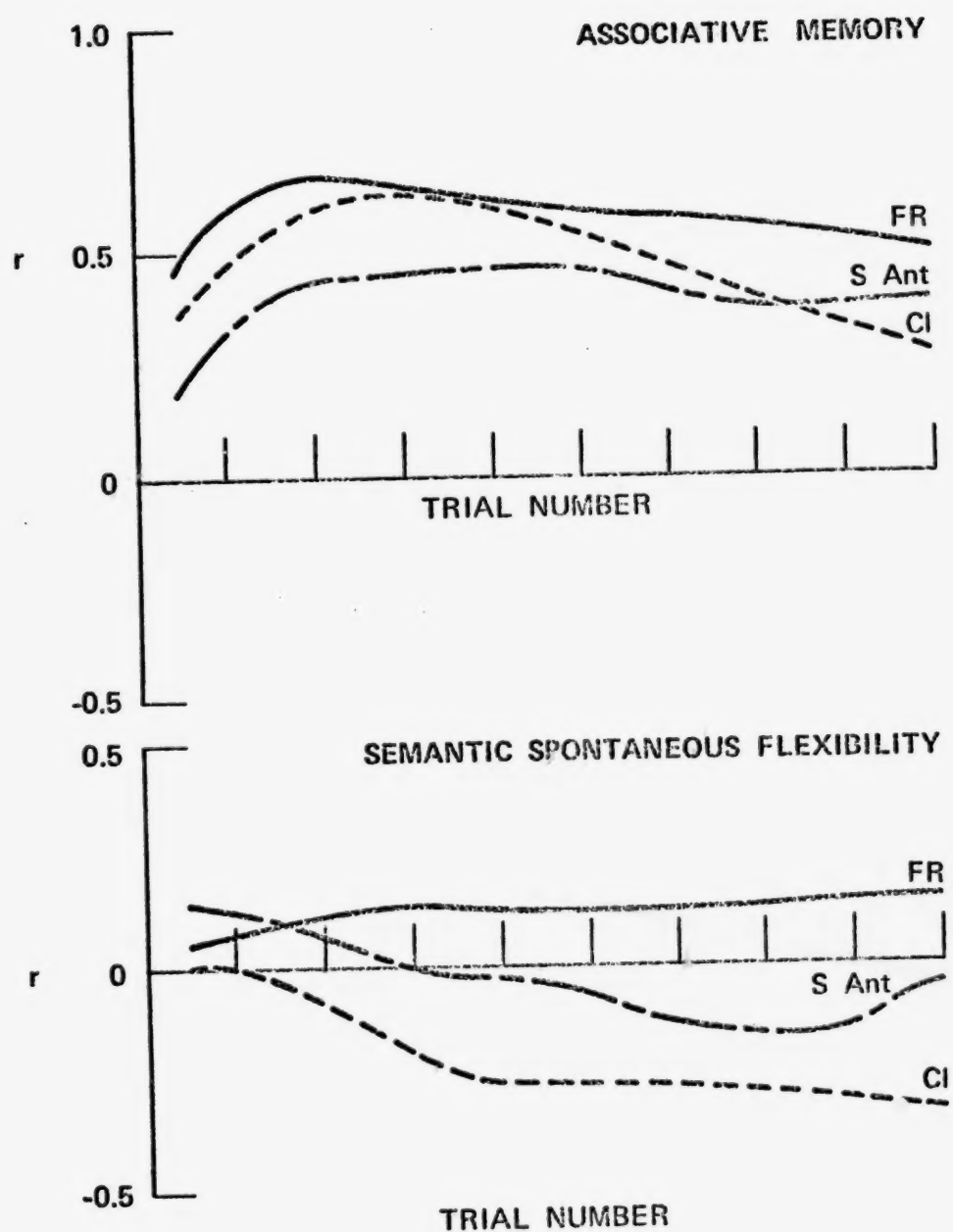


Figure 13. Correlations of ability tests with learning trial performance for three treatments (FR = free recall, S Ant = serial anticipation, and Cl = cluster).

under free recall on trials 3-9 and on two scattered later trials, under clusters treat only on trials 9 and later, and never under serial anticipation. Abilities correlated only modestly with choice of strategy within treatment. The interactions were sufficiently unprecedented that no satisfactory interpretation is possible.

The study unmistakably demonstrates, however, the value of a conception of learning as an active process in which the subject applies a strategy which may or may not fit the task requirements and his abilities. An obvious extension of treatment modifications is to teach strategies, as well as to alter task demands.

While this study does find interactions for specialized abilities, it is a rote-learning study and therefore not of the type treated in the next section.

A curriculum evaluation study.

We would like to include a major section summarizing comparative curriculum studies in which an effort was made to identify what kind of pupil profits best from each of two alternative approaches to a substantial body of subject matter organized in two distinct ways. But such studies appear to be virtually nonexistent.

A fine study by Herron (1966) shows what might be done. He went to four high schools where chemistry was being offered by the conventional curriculum or the CHEM Study curriculum, and collected pretest and posttest data. His achievement measure was planned to fit both courses and to provide sufficient preamble to items requiring special knowledge so as to compensate for any failure of the course to emphasize that knowledge. He tried to prepare items at several levels of the Bloom Taxonomy and analyzed the data within levels. He used the Iowa Test of Educational Development as a general aptitude measure.

The analysis showed no interaction for tests of knowledge, comprehension, application, synthesis, and evaluation, nor for the Watson-Glaser critical thinking test. But there was a significant interaction of some magnitude involving the "analysis" items, which call for reasonably subtle reasoning about the chemical content. The CHEM pupils did better if they were in the higher ranges of v:ed and did relatively worse when they were low in v:ed.

This is consistent with some other hints that complex instruction serves the capable students, and simple instruction the weaker ones. But the absence of relationship for so many other categories makes the effect seem peculiarly specific. Since assigning items to Taxonomy categories is always arguable, Herron might have extracted further information by an analysis at the item level. If he were to identify the items that interact with treatment and attempt to categorize them, he might find some rubric only loosely related to the Bloom system that would account better for the results.

Teacher expectancy and aptitude "change"

Another activity of the project has sought to reanalyze data reported by Rosenthal and Jacobson (1968). This activity deserves mention in our final report, though its progress was arrested in an early stage by the termination of the project contract. The Rosenthal-Jacobson work relates to project concerns in several ways. First, its general conclusion that general ability can be dramatically modified by simple manipulation of teacher expectancy has been used by some to condemn the collection and use of aptitude information in schools. This challenges assumptions on which our project is based. Second, if teacher expectancy is such an important treatment variable, it should be investigated thoroughly in the context of ATI work since it is likely to be a treatment variable that affects different pupils in different ways. Further, in the process of conducting and reporting their research, Rosenthal and Jacobson fell into many of the methodological faults we have come to see as major shortcomings in educational research generally and ATI research in particular. Thus, a review of the Rosenthal-Jacobson book was completed (see Snow, 1969) and a reanalysis of the data was planned. Prof. Janet Elashoff has collaborated in these efforts. Work has begun, but it would be premature to report results at this time. It is expected that the reanalysis will be completed under other auspices, after some delay.

F. Specialized Abilities with Their Possible Interactions with Treatment

Content variables

Since the early work on differential abilities, they have been seen as potential bases for allocating students to different kinds of instruction. No one doubts the relevance of mathematical training for advanced work in, for example, engineering; this and other guidance decisions make use of ATI of a sort. But our interest is in the possibility of achieving the same educational outcomes by choice of methods. There have been many statements as to the possible relevance of ATI of this kind. In the Feierabend conference report (1960, p. 53, p. 112), for example, Gagné predicted that persons high in spatial ability "should acquire mediating spatial concepts more readily than they do symbolic or verbal ones." He extends this to other aptitudes: "The possession of a high degree of spatial ability should facilitate the learning of spatial concepts; high verbal ability should facilitate the learning of verbal concepts; and high numerical ability should facilitate the learning of symbolic concepts."

Carry (1967) conducted a dissertation comparing geometric-graphical vs. algebraic-analytical presentations using programmed instructional materials in the mathematics of quadratic inequalities. Criterion measures representing both immediate recall and transfer to new problems were obtained for 181 high-school geometry students. Carry hypothesized that spatial visualization would be called for in the graphical treatment and so would predict success in it, more than in the algebraic treatment. He hypothesized also that general reasoning would relate more highly to learning from algebraic than from graphical instruction. The data did not confirm these hypotheses. No interactions were obtained with the recall criterion for either aptitude variable. Significant interaction was detected for the transfer measure, but the low internal consistency of this measure made overall findings suspect. Analyses at the item level showed two of the eight transfer items involved in interactions with aptitude. For both items, the reasoning measure was found predictive of responses in the graphical treatment but not in the analytic treatment. For one item, spatial aptitude also predicted graphic but not analytic achievement. Without confirmation, results such as these are uninterpretable.

Nancy Hamilton (1969; Technical Report Number 5) designed a study to test directly the possible relevance of diagrammatic instruction as a variable that would capitalize on spatial ability. Approximately 200 junior-high-school students were assigned to two presentations of PI in crystallography. One program was entirely presented in words. The pictorial treatment was largely but not entirely diagrammatic; it was obviously necessary to introduce names of crystals, for instance, in verbal form, and key terms such as symmetry also appeared as part of the text. The aptitude measures were a Wide-Range Vocabulary Test and a spatial orientation test devised from Thurstone's Cubes; a Punched-Holes test for the visualization factor was also included. The analysis by means of the general linear hypothesis indicated that a regression equation taking all variables into account, fitted within each treatment, did not produce significantly better prediction than an overall regression equation adjusted for treatment means. Hence significance tests for more specific hypotheses could not be justified. The null hypothesis was accepted.

This study, along with that of Carry, makes it necessary to reconsider the widely credited hypothesis: that verbal ability implies ability to learn from verbal treatments, spatial ability, ability to learn from spatial treatments, and so on. It now appears that a quite different formulation may be appropriate.

What do we mean by "a spatial treatment?" Obviously, we mean a treatment that makes use of diagrams. But a spatial treatment may be designed so as to demand considerable spatial reasoning, or it may be so brilliantly executed that the program serves as a prosthesis for the pupil who has poor spatial ability. That is to say, it does his reasoning for him.

While educators have insufficient experience with spatial instruction to say just how this is to be accomplished, we do have such experience with verbal instruction. We know that we can write text material so as to be entirely explicit, with every idea developed slowly and reiterated in a manner that even the pupil with very low verbal aptitude can comprehend. We could, on the other hand, present the argument more elliptically, leaving the student with responsibility for tracing connections, summarizing, etc. It is not obvious that the more explicit

treatment is best. In the first place, it is likely to take longer; moreover, it is likely to be tedious for the able student. Perhaps still more important psychologically, it is likely that the able student who does the work of organization for himself will learn more than when he is a passive recipient of totally pre-masticated materials. Once this has been said regarding verbal materials, it seems obvious that one can similarly design materials either to capitalize on any other ability the student already has and force him to exercise it, or to make it unnecessary for him to rely on ability that is clearly undeveloped in him. This means that research on aptitude treatment interactions cannot be properly designed in terms of hypotheses about "verbal treatments", "spatial treatments," numerical treatments", etc. The Hamilton data showed remarkably similar correlations for the two treatments. The visualization test (which requires mental rotation of plane figures) correlated 0.456 with the principal test score in one treatment and 0.495 in the other. The spatial test, which requires reasoning about the faces of a cube that are turned away from the viewer, turned out to be unrelated to either treatment, the two correlations being 0.033 and 0.148. The correlations for vocabulary (and therefore presumably for all verbal comprehension) were 0.506 for the pictorial treatment and 0.357 for the purely verbal treatment. While this last difference can scarcely be interpreted in view of the failure of the overall test to reach significance, it is consistent with the remarks which we have just made. The pictorial task provided as minimal a verbal presentation as possible, and therefore offered little help by way of redundancy to the person who is weaker in his verbal development. The verbal treatment was not so lucid as to remove all difficulties for the less verbal subject, but the difference in correlations suggests that it was a step in that direction. Yet this difference is the reverse of what was anticipated in planning the experiment and in other discussions of the relevance of "content" abilities. The reader may be surprised to see a substantial correlation of a spatial ability with a verbal treatment, but it will be remembered that the subject matter here was highly spatial, so that spatial reasoning was no doubt required to visualize and comprehend the material. We suggest that the Hamilton pictorial treatment did not appreciably reduce the demand for

spatial reasoning. Treatments can be designed, we now believe, with various degrees of verbal and spatial demand. The treatment making the least verbal demand, however, very likely would not be the one that had the smallest number of words. The question is what processing the subject has to perform to get the essentials from the instructional materials, rather than how the materials appear on the surface.

A comparison of modern with traditional methods of instruction in high school algebra was carried out by Osburn and Melton (1963) using PMA and DAT variables as predictors. The original analysis was made in terms of correlations and showed several rather puzzling differences. Analysis of the raw data reported in Cronbach and Gleser (1965, p. 176) gives information on regression slope. The importance of this kind of analysis is shown, for example, by the fact that DAT Spelling correlated 0.47 and 0.57 with achievement in the experimental and traditional groups respectively; this is anomalous and inexplicable. But when the slope is calculated, it is 0.068 in one group and 0.069 in the other. The first essential finding was that the experimental treatment produced a significantly higher outcome s.d. There were apparent interactions, with the regression slope for the experimental treatment being greater with spatial and abstract predictors. Slopes for verbal ability were positive, uniform for the two treatments. Thus the new-math approach somehow did capitalize on high spatial ability, but we do not know enough about the instructional method to understand this. The practical benefit from placement on the basis of aptitude would not be large, but if the interaction were understood it could no doubt be enhanced.

Throughout this report, where we comment on the practical value of an interaction, the evaluation is impressionistic. A fully serious evaluation would have to bring in considerations of cost and utility -- how important is it to gain an extra two points on the average? How great a price in inconvenience is one willing to pay to run alternative treatments in a school? There is a middle ground, and we would recommend increasing use of such intermediate analysis as ATI research moves out of the purely exploratory and toward tryouts of serious instructional materials. It is not possible to express results in the form of standard scores or the like, since the within-group s.d.'s differ and since the range of individual differences has no particular signifi-

cance on the utility scale. We suggest that for any dependent variable the following be reported: mean (or median) for each treatment, assuming that 100% of the cases are assigned to it; and mean or median if cases are assigned to whichever treatment they are predicted to do best in. Interpretation is made richer if the total possible score is known. These calculations can be made either from the actual distribution of aptitudes in the sample or from normal distribution assumptions.

Another study that gives a hint of relevance of special abilities is that of Hills (1957). Hills, like Hamilton, employed separate spatial and visualization measures, namely, the boat-prow pictures of Guilford and Zimmerman and a clock-rotation test from the same battery. The former measure was strongly related to success in college mathematics courses for engineers but not in sections for physics majors. The second test, however, correlated better in the physics sections. This study is not readily interpreted because we do not have regression functions for the two kinds of classes and cannot judge how much restriction of range entered, and also because we know nothing about the way the courses were conducted. Replications unfortunately have been lacking.

That fairly complex psychological explanations may lie behind any findings about special abilities is suggested by Ferguson and Maccoby (1966). They studied groups of children with peaks or valleys in their profiles on verbal, numerical, and spatial abilities. These peaks are interpreted as reflections of past learning. It is suggested that the verbal peak marks a "bookworm" type with poor peer relations and conflictful dependency on adults. A numerical peak betokens a masculine syndrome of assertiveness and good interpersonal relations and a space peak is associated with behavior inappropriate to one's sex. If interpretations such as these can be sustained, they suggest both possible alterations of instruction and possible lines of remedial effort.

Incidental mention may be made of another study that used existing mathematics classes, as designs of this sort can be expected to be used in the future even though they give relatively ambiguous results. Guilford, Hoepfner, and Petersen (1965) administered a large battery of aptitude measures to classes at four levels of ninth-grade mathematics from "basic" to "accelerated algebra". The data were not analyzed to test for differences in regression slope, nor were regression equations crossvalidated. The study might have placed more emphasis than it did

on the character of the regressions. The only remark to be made here is a warning that, when assignment is nonrandom, regression equations on raw predictor scores or factors derived from them will be misleading. Where construct interpretations are at issue, some step needs to be taken to form regression equations on the basis of correlations for true aptitude scores. No examples of analyses such as this have been reported and we anticipate that there will be many pitfalls in such attempts.

One large-scale interaction study was conducted by Edgerton (1958) in Navy technical training. One method of instruction, applied in a course for aviation mechanics, was essentially rote; trainees were to memorize what they were told and reproduce it on examinations. In the other method the instructor was directed to ^{present} explanations and urge students to raise questions. There were 150 subjects in each experimental treatment. The PMA Tests of Primary Mental Abilities were used as predictors and objective posttests were used. The alternative treatments were given to different classes, and there was a small pretest difference (not significant except in arithmetic) favoring the "why" group; they were also superior on the posttests, though not by a large amount (75.7 for rote; 77.9 for why). Instructors evidently did not execute the treatments as differently as intended.

The original report emphasizes correlations within treatment groups and their significance. As we have pointed out repeatedly, regression slopes are more relevant. For the purpose of the present report, we have calculated slopes for several ability tests against the final achievement composite in the main study. (There is no indication that separate treatment of achievement part scores would be revealing.) In each pair, the figure for "rote" treatment appears first.

<u>Aptitude measure</u>	<u>Number</u>	<u>Verbal*</u>	<u>Space</u>	<u>Fluency*</u>	<u>Reasoning*</u>	<u>Memory</u>	<u>GCT</u>
<u>s.d. (x10)</u>	169;186	168;160	201;219	154;141	87;89	39;45	66;60
<u>Achievement s.d.</u>	968;905	same	same	same	same	same	same
<u>r with ach**</u>	322;205	614;403	444;376	409;184	609;387	260;230	747;631
<u>Regr. slope**</u>	184;100	353;227	213;155	257;118	678;394	645;462	1096;951

(*Difference between r's reported significant; **decimals omitted)

A further more limited study gave essentially similar results.

It is not feasible for us to determine the significance of the differences in slopes, but it does seem that taking into account differences in achievement s.d. may have raised to significance some interactions that did not show up as significant in the correlations. In any event, it appears that general ability of a verbal sort was more highly correlated with success under the rote treatment than under the more meaningful treatment. The net effect, then, was for the explanations to overcome some of the learning difficulties of duller men.

The authors did make a multiple-regression analysis using twelve aptitude scores as separate predictors (including the parts of GCT). The multiple correlation for all predictors was 0.83 for the rote treatment and 0.70 for the why treatment. We are somewhat suspicious of the weights reported, as when two tests are correlated, both will tend to receive weights lower than their separate relevance would indicate. We therefore do not accept the author's conclusion that Fluency was relevant only to the rote treatment and that Memory interfered in the meaningful treatment. The best conclusion here is that v:ed is a more powerful predictor than the more peripheral abilities. A better analysis for the purpose of studying differential abilities would be to partial out the first principal component of the aptitude tests, and then determine if any of the residuals had significant predictive value.

A multiscore interest test had also been given. Correlations were rarely large, but there were some significant differences between treatments. The rote treatment seemed to get best results from men interested in the kind of content being taught, which is not surprising. Performance in the why treatment did not show any relation to interests, implying that the more meaningful treatment overcame whatever handicap lack of interest entailed.

We come now to the Kropp-Nelson-King (1967) project, a program of work that has much in common with ours. Those investigators launched their work somewhat earlier than we did, and proceeded with wideranging and prodigiously energetic explorations. There has been some interchange between the two projects, and our thinking has been able to profit from their experience. We shall be able to summarize its 200-page report only selectively. It does not seem appropriate to comment

on every study, since many of the studies were preliminary and were superseded by more complex studies in the same area. This group at Florida State University was strongly influenced by Guilford's hypotheses; we, having reexamined more of the Guilford data, have already indicated our reservations about the ultimate worth of that system, and this will be recognized in our interpretation of the findings; but we would not have been so skeptical at the time these studies were initiated.

One line of work had to do with redundancy of reading material. The study reported on p. 57 ff. was an ATI experiment using sixth graders with generally low IQs and presumably low socioeconomic status. Half the subjects read an original textbook version of a science narrative; the others read an especially prepared "redundant" version. Redundancy had been achieved by various kinds of simplification. The redundant version was not, however, repetitive; on the contrary, it was nearly 20 per cent shorter than the original. After pupils completed the story they took a multiple-choice test on the common content. As aptitude measures three PMA tests and a syllogistic test from CTMM were given. The data show a clear effect; the abler students did better on both versions, but had a particularly large advantage on the simplified version. While linear regression lines crossed in the lower part of the range, it is possible that a curvilinear analysis would have been better, since a floor effect may have occurred in this sample. Other aspects of the analysis may be questioned. A single score was obtained by dividing the pupil's achievement by his study time; whether or not it was feasible to use a bivariate dependent variable, it would have been well to inspect the data for achievement and time together to make sure this arbitrary way of combining them did not becloud results. Second, in this and some later studies, interaction was tested by comparing regression slopes for aptitude variables in turn; a multivariate treatment would have been much more satisfactory. By this analysis, only the PMA Reasoning test showed a significant difference, but the trends for all four variables are similar and it seems best to account for the results in terms of a general factor. There is no convincing evidence that this variable was significantly more powerful than the others. (Here and elsewhere, some of our

moderately critical remarks echo comments made by the original investigators. Moreover, they make additional comments on the conduct of ATI research that we have not quoted, so that the reader of this report would do well to consult theirs for the general methodological insight it can provide.)

Substantively, we find this study puzzling, especially as we earlier found "smooth" PI treatments best for weaker pupils. In connection with the Hamilton study also we speculated that increasing verbal redundancy would simplify comprehension by the less able pupil. Just the opposite occurred here. By the cloze test, the revised passage was more redundant even though briefer; this redundancy was achieved by simplifying and removing detail. More research with redundancy as a variable should be carried out, at various ability levels. Repetition might better serve the less able pupil than sheer simplification does.

We skip over a study in the mathematics area that relates to other work on inductive vs. didactic instruction to which we shall turn below. Relevant here is an attempt to introduce a verbal vs. figural distinction in treatments, in a manner essentially like Hamilton's. This gave no significant interaction with PMA verbal and perceptual tests (p. 85). Using two kinds of didactic or inductive treatment does help substantially in examining whether a conclusion can be generalized.

We may also deal summarily with a study of concept attainment using either figural or verbally described stimulus instances. While a very small pilot study suggested that fluid ability (Hidden Figures) was associated with better figural performance and poorer verbal performance, there was no hint of such a relation in the main study. Unreliability of concept attainment scores under the conditions used was a major difficulty.

It is impossible to give an adequate brief description of four studies dealing with the learning of vocabulary by three different methods. One unusual design feature was that students learned different sets of words by each of the instructional methods, in counterbalanced order. There were nine Guilford aptitude measures, and in some studies two or more dependent measures. There are no tests of significance of interactions and the variations in correlations reported are for the most part clearly in the chance range. This is a study where a test

of regression of outcome onto the principal components of the dependent variables would have been an economical strategy. A particularly interesting study presented a new word list on each of ten days, so as to measure trends over time. Unfortunately, attrition left the sample size too small for firm conclusions. There is a hint in the data that under one training procedure memory was a good early predictor and divergent thinking a good late predictor.

We come now to the final series of studies, which profits from the earlier experience. The reader will have noted the unusual readiness of these investigators to do pilot work and to replicate their studies; here it proves especially valuable in verifying a generalization. The basic distinction on which the study rests is the Guilford distinction between semantic (meaningful verbal) tasks and symbolic tasks where formal elements must be manipulated. It was supposed that mathematical materials could be presented in either mode and that relevant Guilford tests would interact with the treatments. The treatments dealt with vector multiplication and the taking of derivatives. An example of the variation is the contrast between these corresponding parts of frames in the two treatments:

Semantic

Definition
The product of two vectors, when the two vectors are expressed as ordered pairs of real numbers, is defined by the following three steps:

The steps are stated verbally in the first case and algebraically in the second. The second treatment resembles the one Carry called analytic. College psychology students worked through one or the other program during a three-hour period. There were 10 aptitude tests and three posttests. The authors' basic analysis was the computation of regression slope for one aptitude at a time within each treatment. Some significance tests for differences in slope were calculated.

We list here certain key differences, rearranged. In general, these refer to total achievement scores rather than to subcriteria. It will be recalled that in Guilford's system semantic tests are coded -M- (e.g., CMR) and symbolic tests -S- (e.g., CSR).

Symbolic

Definition
The product of two vectors $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ and $\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$, written $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} * \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$ is defined by:

Tests for which slope under semantic treatment was at least 1.5 times that under the symbolic treatment (or clearly greater and opposite in sign):

Study 1	Study 2	Study 3	
CMC	CMC		
CMR	CMR	CMR	
NMR	NMR	NMR	Semantic tests
NMT	NMT		
	NMI	NMI	
NST	NST	NST	
NSI		NSI	Symbolic tests
	CSR		
	NSR	NSR	
		CSC	

Tests for which slope under semantic treatment was no more than 0.67 times that under the symbolic treatment (or clearly less, and opposite in sign):

Study 1	Study 2	Study 3	
NMI			Semantic
			Symbolic
	NSI		

Study 1 and study 2 were essentially similar and conducted on similar kinds of subjects. The findings are very similar, the more so when we realize that the aptitude tests are fairly unreliable and the sample size modest. It is evident that the semantic treatment is more dependent on measured aptitudes. Where any test appeared to relate more strongly to outcome from symbolic treatment, replication gave a contradictory result. Here, as in Guilford data, the two categories of tests do not cluster particularly, so that one could not expect to establish a strong generalization regarding category differences. While a few intercorrelations were negative, there were enough positive intercorrelations to suggest that it would be well to extract one or two main factors, and calculate regression slopes on these more reliable measures. It seems obvious that a fairly general factor would have included most tests of both kinds and would have entered into a significant interaction, with

steeper slope for the semantic treatment. There is no obvious indication that a second factor would have differential validity or, indeed, any validity in the population -- but we cannot be sure of this.

Study 3 was conducted on 177 tenth graders, using the same materials. This time the instruction extended over three days, one hour per day. The program was clearly too difficult for them, at least under spaced conditions. The authors are inclined to see the Study 3 interactions as inconsistent with the other two studies, but that is the case only if one takes very seriously the multifactorial breakdown of the data. The first factor we suggested using would almost certainly produce a significant result in this study also, even though certain components such as CMC has a steeper slope for the symbolic treatment that falls only a bit short of our 0.67 criterion. Nonetheless, recognizing the treachery of multivariate work with short tests and small samples, we consider it likely that a reduced-rank regression analysis would tell about the same story in the three studies.

What we have, then, is one of the few solid indications that it is possible to design a genuine group instructional treatment that serves test-weak pupils better than conventional verbal treatment does.

This conclusion is not a victory for differential testing. It is but a small variant on the theme that general ability predicts under some treatments better than others. In this case the two treatments were more or less equal in difficulty, as judged by test means, but the symbolic treatment was very likely a good deal more novel to the subjects. (This would be a good study to carry out with a learning-to-learn design.) One might argue that the semantic treatment was more redundant, and count this a confirmation of the earlier study on redundancy. We would prefer to leave this nicely established result as a matter for further research to explain.

Among the general remarks of the Florida group, two are especially worth our repeating. One has to do with the difficulty of writing posttest items that are equally fair to both treatment groups, in most of the studies. This difficulty was encountered by Hamilton also. But the results of both the Hamilton and Florida studies are somewhat reassuring. We do not find significant evidence that different kinds

of posttests give different reports on ATI. This is not likely to be a universal finding, however, and some attempt to provide test items suited to both groups is wise. Second, there is a problem of equating study time, within and between groups. The person who spends more time on an instructional program may learn either more or less than the one who spends less time; this holds a fortiori where different programs are concerned. No prescription for equating can be offered, though one alternative worth considering is to provide so much by way of review and extra examples that all students have something to keep them busy throughout a fixed study period, uniform for all subjects. This is realistic enough, as classroom schedules go. Where time is variable, it should be recorded and somehow taken into account in examining the data; it is an uncontrolled treatment variable or a mediating variable rather than an aptitude or an outcome, however.

Attention may also be called to an unpublished dissertation by Behr (1967) which followed the studies mentioned above. Here again the center of interest was variables from the Guilford system. Mathematical material was presented either in "figural" or "verbal" form, with the supposition that figural tests would predict the former and semantic tests the latter. Over 200 elementary teachers-in-training were asked to study numerical operations in modulus-seven arithmetic by one of two programs. Seven outcome measures were crossed with fourteen separate tests, thus producing a breeding ground for chance relationships. A properly powerful multivariate analysis, using the principal dimension or dimensions in the outcome data, would be required to interpret the study. Behr himself calculated regression slopes for one Guilford measure at a time, and emphasized certain scattered tests where regression slopes seemed to differ from treatment to treatment. We have made use of these slopes (given in the appendix to his dissertation) to examine a simple hypothesis in a crude way. Taking all figural tests (e.g., CFU, MFS) we summed the regression slopes for the main outcome (Total score on learning test); this was repeated for each treatment and for semantic as well as figural tests, with these resultant slopes:

	Figural tests	Semantic tests
Figural treatment	7.45	4.28
Semantic treatment	9.21	7.45

The numbers should not be compared from column to column, as predictor variances differ arbitrarily. It is evident that a general factor predicts better in the semantic (verbal) treatment; essentially the same result is found for a retention measure. This is consistent with the finding of Kropp et al. Possibly a multivariate analysis would show something further; the tests where the slope difference is most striking (higher in ^{the} verbal treatment) are CFU, MFU, CMU, and MMR.

Apparently there was a failure to find ATI in a doctoral study by Lim (1968) carried out under the direction of Carroll. We say "apparently" because within treatment correlations and regressions are not mentioned; but analysis of covariance was used, and the technical quality of the study is such that we are sure the basic assumption of homogeneous regressions must have been tested. Available as covariates were elementary version of Modern Language Aptitude Test, IQ, and French grade. Treatment variables were presence of pictures or English sentences corresponding to the Malay sentence being practiced, a prompting sequence vs. a confirmation sequence, a deductive vs. an inductive organization, and finally a comparison of structural and transformational-grammar approaches. The main effects favored the confirmation method and the use of English sentences rather than pictures.

Having introduced this section with hypotheses offered by Gagné in 1960, it is appropriate to finish with a recent study that brings Gagné's thinking on the matter down to date. Gagné and Gropper (1965) undertook an ingenious study of the effect of pictorial instruction that has several excellent design features. Eighth-graders were given aptitude tests and pretests. They were then taught by two self-paced programs on content in mechanics; this served as a "tuning" period and guaranteed acquaintance with many basic concepts to be used in the study of principles of mechanical advantage (e.g., in levers). There was then a self-paced instructional program on mechanical advantage. The control group had this alone. The visual-treatment group had an introductory demonstration, presented in a fixed-pace manner via television. Evidently, students viewed this and then as they worked through the self-paced lessons they saw the same or new visual

explanations. The verbal-treatment group was treated similarly save that in place of pictorial demonstrations they had essentially the same content in verbal form. It is not at all clear how rate of learning on the self-paced program was timed, in view of the "interspersing" of the treatment material. The essential idea here was to use learning rate on new material as a dependent variable; it was hypothesized that the demonstrations (advance organizers?) would be helpful in the new learning and that the effects would interact with verbal and spatial aptitude (DAT). Abstract (DAT) and v:ed (Otis IQ) were also measured but were not expected to interact. As additional outcomes, there were an achievement test and a delayed achievement test a month later.

The analysis emphasizes correlations rather than regressions and does not test for significance/^{of} differences in r across treatments. We have been able to calculate regression slopes, given in Table 4. It is worthwhile first to note that the supplementary treatments were somewhat helpful in promoting achievement and retention, but that their effects on learning rate were negligible. The learning-rate variable is open to question, despite its theoretical importance, since there was no control to guarantee that students would master the material as they proceeded. It is hard to judge which differences in regression slope might be significant, and data are not available for calculating a proper significance test. The reader is warned not to dismiss numerically small values as trivial; correlations ranged as high as 0.48, which implies a considerable effect. There is a weak tendency for the regressions on Space to be steeper in the visual treatment. Learning rate is sharply related to IQ in the visual treatment and little related in the verbal treatment. But the effect is just the opposite in the test scores. If we pool the first two treatments and contrast them with the control, we see a striking tendency of achievement test and delayed test to be more strongly related to IQ and Verbal ability in the control group. This patterning is not brought out in the Gagné-Gropper analysis of correlations.

TABLE 4

Regression Slopes of Outcome on Aptitude
in Gagné-Gropper Study

	<u>Learning rate</u>	<u>Posttest</u>	<u>Delayed test</u>
Regressions on Space Reasoning			
Visual	+.10	.16	.14
Verbal	-.06	.14	.05
Control	-.04	.27	.24
Regressions on Abstract Reasoning			
Visual	+.01	.46	.48
Verbal	+.03	.46	.26
Control	-.06	.50	.47
Regressions on Verbal Reasoning			
Visual	+.53	.69	.59
Verbal	+.36	.52	.63
Control	+.16	.82	.84
Regressions on Otis IQ			
Visual	+.57	.20	.19
Verbal	+.17	.33	.28
Control	-.21	.80	.66

The results do not at all support the initial hypothesis of steep slopes for the visual-space combination and the verbal-verbal combination. The authors wisely refuse to interpret the scattered evidence of special-ability interactions. Some of their comments will be of considerable interest to our readers, especially as the study has rarely or never been cited and hence these pertinent comments have not been brought to the attention of the investigators. It is necessary to paraphrase rather than to quote the discussion. They decide that abilities of the sort measured by most tests are too general, and that if ability to learn from visual materials depends upon specific abilities, they will be those directly related to processing information in the task. Some such abilities are ability to abstract a class concept from exemplars, ability to discriminate visual objects by multiple cues, ability to resist interference in the face of frequent reversal of cues in object identification, ability to code unfamiliar figures for retention, and ability to identify correct verbal statements of principles from visually presented specific examples. Abilities like these are not measured by present tests, but they are capable of being measured. Indeed, such elements have probably been discarded from past aptitude tests just because of their specificity. Gagné and Gropper ^{advocate} /analysis of the particular class of tasks to be used in learning, and using these to conceptualize the relevant aptitudes. In sum, five years after Gagné made his recommendation that content aptitudes be brought into interactional studies, he decided that a task analysis of what the test, and the instruction, call for is ^a more appropriate source of hypotheses about ATI.

By way of summary of these studies, we offer the table below. It is evident that a priori hypotheses often fare badly. There are several interactions but it is not at all clear why they happened. There are no successes in planning treatments to capitalize on spatial or figural ability, though two studies with uncontrolled treatments reported interactions of outcome with these special aptitudes. The last of the Kropp studies is far more impressive than the others because of its greater control; but hypotheses regarding content-differentiated abilities did not work out.

	Treatment Contrasts	Aptitudes	Interaction	Remarks
Carry	(A) Geometric (B) Analytic	Space, reasoning	Dubious	Two transfer items were predicted by reasoning, in (A) group; contrary to hypothesis.
Hamilton	(A) Spatial (B) Verbal	Space, visualiza- tion, verbal	No	
Gropper	(A) Visual (B) Verbal	General	Dubious	(B) best for able students in some comparisons.
Osburn- Melton	(A) "Modern" (B) "Traditional"	Various	Yes	(A) best for high abstract and high spatial.
Hills	(A) Math for engineers (B) Math for physicists	Space	Yes	(A) best with high space, (B) with high visualization.
Edgerton	(A) Rote (B) Meaningful	Various		(B) better with high gen- eral ability; no differ- ential-aptitude interaction.
Kropp, et al.	(A) Traditional text (B) Simplified	Various	Yes	(B) better with high gen- eral ability; no differ- ential-aptitude interaction.
Kropp, et al.	(A) Figural displays (B) Verbal displays	Various	No	Two studies of this type
Kropp, et al.	(A) Symbolic analytic (B) Semantic	Various	Yes	Replicated in three studies: (B) more strongly dependent on a fairly general ability, no differential-aptitude interaction.
Behr	(A) Figural (B) Verbal	Various	No	(B) more strongly dependent on a fairly general ability, no differential-aptitude interaction.
Gagne- Gropper	(A) Visual (B) Verbal (C) Control	Various	No	Verbal and general related more to success in (C). No dependable separation of (A) and (B).

Discovery or induction

In the literature on "learning by discovery" it has been suggested that such an approach might serve pupils who have done poorly under conventional methods. One study of this kind was conducted by Becker (1967) using Gagné's instructional programs on number-series formulas. Becker hypothesized that among high-school algebra students, those high on verbal tests and low/^{on} mathematical reasoning would achieve better in the "didactic" treatment, and low-verbal, high-quantitative students would do best with "discovery". His results were negative. (Here as in ^{some} other studies we have reviewed, a pilot study had shown a seemingly significant effect.) Regression coefficients for the aptitude variables did not differ significantly between treatments on any of these criteria. Both aptitudes were positively related to achievement regardless of instructional method. Tanner (1968) programmed principles of mechanics for ninth graders. His three treatments were 1) expository-deductive, in which students read rule and principle statements before working examples, 2) discovery-inductive, in which principles were left unstated but a planned order of guiding examples was followed, and 3) unsequenced-discovery, where order of examples was randomized. A mechanical reasoning test showed no interaction. Again, a general pretest on the content correlated positively with all criteria in all treatments with no interaction. Tanner incidentally obtained disordinal interactions with sex for both comprehension ($p < .05$) and lateral transfer ($p < .01$) criteria. "Expository" methods were best for boys while "discovery" produced better performance for girls. An analysis of ATI within sex would have been a useful exploration.

Peters devised three alternative training methods to train conservation of number in kindergarten children. In one treatment -- "non-cued discovery" -- the child was shown transformations of block arrangements where no specific cues to the relevance of the number dimension are available. A second treatment -- "visual-cue guided" -- adds a color-code and a number-code to the blocks. A third treatment is "verbal didactic", adding verbal statements of the conservation rule to the first condition. Interaction ~~was~~ hypothesized between treatment and a combination of learner variables representing language

comprehension and analytic style (Kagan et al., 1964). For children high on language comprehension and low on analytic style, the verbal treatment was expected to be superior; the visual-cue guided treatment was planned to be especially appropriate for low-language high-analytic children. Results did not confirm the hypothesis, though the direction of differences was as expected.

Incidental mention may be given to a study of "laboratory" instruction, not necessarily using discovery. Bush and others (1965) carried out a potentially informative experiment but obscured the results by analyzing raw gains, and using a difference score as an aptitude variable. There was a repeated measures design, each subject rotating through five treatments. If the conclusions drawn can be trusted, students with Mathematics-Fundamentals-higher-than-Vocabulary profited more from individual laboratory instruction, and students with the reverse pattern did best under verbal instruction. No significant effects for WAIS and ACQT scores were reported. The correct analysis would test homogeneity of posttest regression slope, using the basic aptitude scores and the pretest as predictors in a single equation for each treatment.

Reference was made earlier to the Kropp-Nelson-King study in which there were four treatments: verbal inductive, verbal deductive, figural inductive, and figural deductive. The analysis with four treatment groups was not readily interpreted even though there were 100 cases per treatment. Here again, a reduced-rank analysis might have been the most informative way to push aside overdifferentiated, meaningless information. The solution actually followed was to consolidate the two deductive and two inductive treatments. Basically, the "deductive" treatment was a textual didactic presentation of a definition followed by examples. The pupil was a passive reader, any deduction being done by the instructor. In the inductive treatment pupils were given examples of a concept and asked to think of other examples. Induction of a rule was not explicitly required. The instruction, covering a topic in set theory, was given to fourth and sixth graders; the instruction and posttest occupied only one session. The inductive treatments gave somewhat better results, and were somewhat less well predicted by aptitude measures. (R^2 with six aptitudes 0.30 - 0.36 for inductive, 0.39 - 0.44 for deductive. The similarity of

these values argues against the presence of a practically useful interaction, as different weights were fitted in each treatment.)

ATI analysis was made only for three aptitudes originally hypothesized to relate to the I-D distinction. A significant interaction was found which is most easily summarized as follows: a verbal syllogistic test contributed most to prediction of outcome under D and inductive reasoning tests (figure and word grouping) to outcome under I. This result, on closer inspection, appears to derive from some discrepant standard deviations on the aptitude tests (despite random assignment). We have recalculated slopes within the original treatments, with the following results:

	Deductive V	Deductive F	Inductive V	Inductive F
Syllogisms	.64	.48	.42	.32
Figure grouping	.28	.03	.32	.32
Word grouping	.20	.45	.38	.50

With the exception of the value of 0.45 in the last row, these results are consistent with the combined analysis. Some differences are so small, however, that one hesitates to put weight on the conclusion. Maybe we have here the rare case where outcomes depend at least weakly on differential abilities. But one would want the study confirmed by instruction extending longer in time and allowing more realistic provision both for the pupil's learning to learn and for his responding actively during instruction.

A kind of guided-vs-pure-discovery comparison was involved in the concept-attainment study of Dunham and Bunderson (1968). High school students worked on nine difficult problems; one group had a general orientation and the other group was given two highly complex "principles" to use. The treatment was basically ineffective; it did not raise the mean appreciably, and both groups did rather badly on the problems. Twelve aptitude measures were given and, with a judgment rare in the studies we review, the authors performed a principal components analysis to reduce the set to six factors. Unfortunately, the results of the ATI comparison are then reported in terms of factor loadings (correlations) rather than regression slopes. (s.d.'s are not reported for outcomes under the two treatments. Ignoring s.d. distorts results.)

The analysis also has the defect of treating the nine problems separately, thus using highly unreliable scores. To obtain a rough indication of what the data actually show, we added factor loadings across problems and conclude that (1) a quasi-general "induction" factor is substantially related to success under both treatments, as is usual in concept-attainment; (2) reasoning ability relates much more strongly in the principle group; (3) other tests, notably memory, correlate more strongly with success in the no-principle group. We are hesitant to interpret the differential data strongly because it seems unlikely that the subjects understood the principles being presented. It would have been worthwhile to measure this understanding and to introduce the measure into the analysis of the principle group.

A summary tabulation for the studies in this category is scarcely worthwhile. Results were clearly negative or highly suspect in nearly all the studies.

For all that has been said about "divergent thinking" as a distinctive and useful type of ability on which schooling could capitalize, we found little evidence on ATI in this area. Hutchinson (1963) completed a doctoral dissertation under Calvin Taylor in which four classes were taught social studies by a method in which the teacher elicits from pupils more independent thinking of evaluative, convergent, and divergent types. The same teachers taught control classes in a didactic mode of giving information and eliciting recall. There were pretest and posttest data for a unit on transportation and communication. CTMM mental age was available and used as a basis for matching and analysis (much preferable to IQ). Seven pretests on divergent thinking were collected but regrettably these data are not given in usable form.

The author is concerned with interaction because he regards the conventional school as suited to one type of "high IQ" student, and hopes to encourage teaching that another group of pupils will profit from. His analysis commits at least two errors -- use of raw gain

scores and analysis of correlations rather than regressions. The mean difference between treatments was fairly small in each class, and may be ignored. It will be useful to compare the author's report with our own conclusions from the data. He gives the following correlations between mental age and the gain score:

	Control	Experimental
Teacher A	0.51	-0.02
Teacher B	0.41	0.19
Teacher C	0.44	-0.16
Teacher D	-0.50	-0.17

The author concludes that there is a significant interaction in classrooms A, B, and C.

The most interesting analysis for this study would employ three pretest variables: mental age, score on divergent-thinking tests, and score on content pretest. It would be particularly suitable to factor analyze (ideally, with reliabilities in the diagonals) and determine regression weights of the posttest on the factors in turn. This we have been unable to do. In the first place the author does not report any of the correlations we would use. Second, the necessity to present this report at this time prevents us from making a full analysis of the data we have. We first correlate MA with posttest, for the interest it may have:

	Control	Experimental
Teacher A	0.64	0.56
Teacher B	0.54	0.63
Teacher C	0.72	0.20
Teacher D	0.39	0.47

Here there appears to be an interaction for C, and presumably none for the other three teachers. Moving on^{to} the raw-score regression slopes we have:

	Control	Experimental
Teacher A	0.240	0.165
Teacher B	.220	.204
Teacher C	.174	.040
Teacher D	.094	.115

The conclusion (not tested for significance) is that there is interaction

for teachers A and C and not for B and D. For Teacher A we can illustrate a multiple-regression solution. The weights are

	Mental age	Pretest
Control group	0.19	0.67
Experimental group	0.12	0.62

(No interpretation should be placed on the magnitude of numbers in the two columns, as this is dependent on the scoring scale.) There was no difference in the regressions on the pretest; ^{there was} a difference on mental age like that determined before.

This study with its four replications suggests a general remark. There ought to be studies calculating regression slopes for outcome on aptitude for classes of various teachers wherever a common outcome measure is appropriate. It is reasonable to suppose that there will be teachers whose pupils' final attainment correlates highly with pretest scores, and others for whom there is only a weak correlation. These two kinds of teachers perform different social functions. The former conserves and enhances already apparent talent -- the second overcomes deficiencies. The important question is, what do these two teachers do differently in day-by-day class management? Unfortunately, although data suitable for such analysis must have been collected thousands of times, we find no record that such a study has been made..

On the general topic of divergent thinking, we may note one more report. Ripple and O'Reilly (1967) evidently were unable to establish an interaction of programmed (vs. conventional) instruction with general ability, divergent thinking ability, or anxiety. (Abstract only seen.)

G. Interactions in the personality domain

The whole conception of personality as a vehicle for psychological theory and practice is in a state of flux. Despite the extended efforts of psychologists pursuing diverse traditions, it appears that each of the methods of testing and conceptualizing personality that has been exploited during the past two decades is open to serious criticism. It is hard to see that any one of the several lines of effort -- empirically-keyed questionnaires, factor analytically-keyed questionnaires, measures of response style, or global assessment -- has moved forward during this period. The deficiencies of the methods and the associated conceptualizations has become increasingly obvious. This is underlined by the appearance of independent books by Mischel (1968) and Peterson (1968) in which the attempt to conceive of personality in terms of "traits" is declared bankrupt. These authors emphasize two things: (1) Any attempt to characterize a person as "anxious", for example, predicts his behavior in any single situation very poorly, because the specific characteristics of the situation do much to determine how he will respond. (2) Even though a person is inclined to respond in a particular way, he is capable of learning a required role -- suppressing overt indicators of anxiety, demonstrating "dominance", or otherwise rising to the occasion. This criticism is justly directed against the use of test scores as a means of planning therapy or as a means of selection. Since there are statistical tendencies for "anxious" persons to display the usual symptoms over a wider range of circumstances than the "nonanxious," there is still a case to be made for trait-oriented scales in research aimed toward theory. It is increasingly clear, however, that to define a trait simply in terms of the responses the person is expected to display is not a very powerful way to search for a theory.

What a person can be expected to do will vary from one class of situations to another, according to these authors; it follows that a sensible measuring instrument will define a reasonably broad but not universal class of situations and inquire about reactions in those circumstances. Thus Endler, Hunt, and Rosenstein (1962) demonstrated that quite different persons turn out to be "anxious" depending upon whether the eliciting situation is an academic threat, a social demand, or a physical threat. One can begin to visualize a

hierarchical conception of traits such as anxiety, in which general anxiety is subdivided along the lines of broad stimulus classes such as those just cited, and these are further subdivided where a fine analysis is useful.

The recommendation of Peterson and of Mischel regarding the practical handling of the individuals lays great stress on ATI. But instead of expecting treatments to covary with broadly-defined personality traits, they believe that it will be necessary to identify the specific situations, significant others, roles, and emotional reactions that describe the individual's response to his own environment, and then by either cognitive or conditioning techniques to work directly to modify those response tendencies. This differentiates treatment according to information about the individual, but it makes no place for generalizations about types of individuals, i.e., for ATI theory.

Another vigorous attack on personality testing has been mounted by those who consider such tasks to be invasions of privacy. Insofar as the objection is to obnoxious questions, it is irrelevant to us here. But another aspect of the discussion is the objection to forcing the individual to "testify against himself", i.e., to give information that may deprive him of a job or admission to an educational program. To force him to reveal his inner feelings at the risk of a penalty is unacceptable; it is equally unacceptable to invite him to lie to the inquirer. Much of this objection vanishes when we contemplate using personality variables as a basis for choosing the kind of treatment the individual will experience rather than to select or reject him. If one can use personality dimensions to suggest the kind of instruction a student will respond successfully to, the testing is wholly in his benefit and he should have no motivation to give false answers. There still must be some provision to protect him from unduly invasive questions, but the test given for placement purposes is much less an adversary proceeding than the test given for selection.

It is impossible to collate results of studies of personality, even though a large number of them that have offered some evidence on the presence or absence of ATI. The difficulty is that investigators rarely use the same measures, often interpret the same measure in terms of different constructs, and almost never employ two operational indicators of the same construct, as is necessary if one is to defend one interpretation as more plausible than the competing hypotheses that are always available.

It will be necessary for us to proceed with a rather crude organizing scheme in order to bring a reasonable sample of the papers together where they can be compared. Any synthesis must be loose and impressionistic, as the data are essentially noncomparable.

Fearfulness

There is a broad complex of variables involving anxiety, confidence, "neuroticism", compulsiveness, and other scores. While the various measures differ in their surface content and are not necessarily well-correlated, one can think of a syndrome of self-deprecation, expecting to fail, seeing the environment as threatening, and so on. The fearful person will adopt various strategies for coping with these threats: withdrawal, denial, and compulsive self-control are common, but there are also compensatory mechanisms. At the opposite extreme, persons are usually characterized as secure and confident. They may or may not be energetic and ambitious; confidence can go with passivity.

The most common type of investigation in this field has selected a group of Highs and a group of Lows on the Taylor Manifest Anxiety Scale or the like. The groups may be separated at the median of the total sample, or they may be chosen from the extreme details of the distribution. The persons are then exposed to one of two treatments, and some outcome is measured. We may anticipate many of the results to be reported by saying that the data tend to be consistent with a curvilinear relationship, such that at each extreme of the scale outcomes are poor, and at some intermediate value outcome rises to an optimum. The optimum for one treatment comes at a different point on the anxiety scale than for another. The treatment that works better at relatively high anxiety levels is likely to be less

stressful in some sense. This kind of hypothesis has been present in psychology since the 1908 work of Yerkes and Dodson on habit formation in the presence of varying degrees of electric shock. The general conception of such findings is that when the organism is already aroused, additional stress or stimulation is likely to be detrimental; when arousal is below some optimal value, a treatment that heightens arousal is likely to be beneficial. Then for any individual there is some best level of stress-producing stimulation (in the context of a given criterion performance to be mastered). And for any given environment, some subjects are closer to the optimum level of arousal than others. With arch-shaped regression lines in mind, we cannot be content with the prototypic experiment that reports mean outcomes for high and low groups, thus giving only two points on the arch. What then looks like a linear regression may be one leg of the arch. What looks like an absence of relation may be comparing points on opposite legs. Moreover, to contrast "high" and "low" groups is essentially meaningless. "High" has different consequences in different parts of the range and no generalization is possible. Any true consolidation in this field will require a reporting of results in greater detail than has been done in the past, and in such a way that outcomes can be projected onto a reproducible scale. The easiest way to achieve this would be to standardize upon a limited number of intercalibrated measuring instruments, but nothing of this sort has been attempted.

"Structure" as a treatment variable. We begin this summary with one of the most imaginative and educationally relevant investigations of ATI, though one that is not too convincing. This study (Grimes & Allinsmith, 1961) was conducted on an a posteriori basis to test the hypothesis that whether a child responds best to phonics or to a less-structured type of instruction depends on his personality. Instead of experimenting, they went to two school systems where different methods of instruction had been in use and

tested children in the third grade. Underachievers, normal achievers, and over achievers were selected on the basis of the regression of achievement measure on a mental test score. The anxiety of these children was measured with the children's version of the Taylor scale. Compulsivity was measured by an interview regarding the child's typical behavior. These two personality variables were uncorrelated. The dependent variable was the deviation of the pupils' reading achievement from the achievement that would be predicted from his Wechsler IQ. While the Wechsler test does not depend directly upon reading, it is highly likely that scores are in part a reflection of the success of the pupil's preceding years of school experience. The use of discrepancy scores is less satisfactory than the use of a bivariate dependent variable would have been.

The school systems were not entirely comparable. In the schools where phonics were used the classrooms were characterized as "more authoritarian and cold". Moreover, although an attempt was made to select schools of similar social-class background, one system was an industrial city and the other a suburb, which makes equating impossible. The essential finding is presented in Figure 14. There were distinctly significant interactions between personality and instruction, even though the structured method of instruction apparently produced better overall results. Apparently structure was beneficial to the high-anxious, high-compulsive child. On the other hand, it was not better than the unstructured method for the pupil who is neither compulsive nor anxious. The various flaws in the design make us hesitant to generalize; it seems quite possible that some other school system could use unstructured methods differently, and achieve superior results.

Although the data are weak, it is striking that the pupils who do best in the structured program appear to be those who are compulsive and anxious. Those who are not compulsive do relatively badly. In the unstructured school, the combination most prognostic of success is high compulsivity with low anxiety; the reverse combination is unfavorable. The suggestion is strong that providing structure enables pupils to use compulsivity and anxiety constructively; the child who worries about what he should do to keep out of trouble can easily see what to concentrate on. The unstructured program, on the other hand, apparently favors the child who is emotionally free to provide his own structure, and is disturbing the child

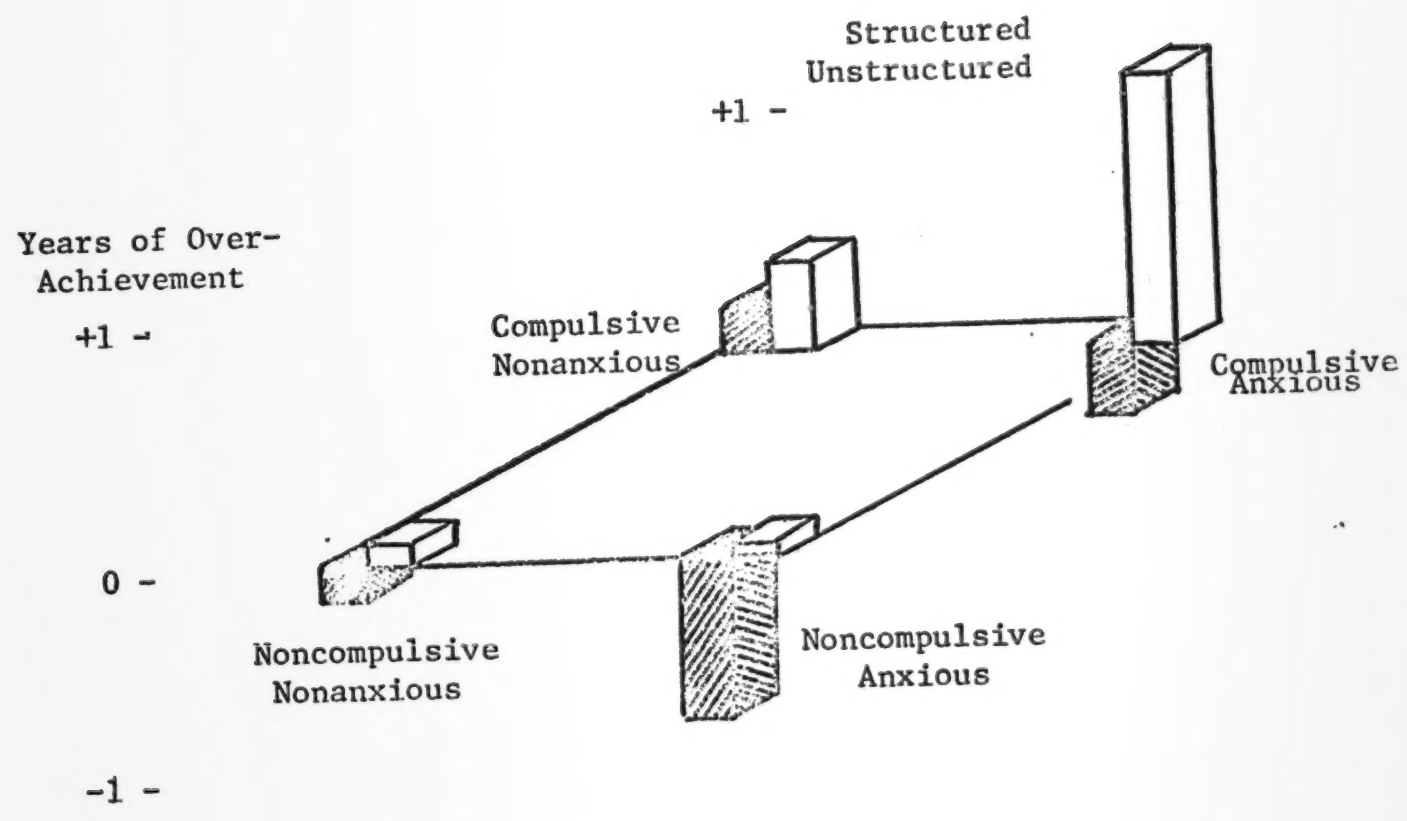


Fig. 14. Results from Grimes-Allinsmith Study

who feels under tension but does not have a systematic way of coping with his environment. One difficulty with this type of interpretation is that personality manifestations shown in the third grade are likely to be in part the result of two years of school experience. It could be said, for example, that the student who has succeeded in the structured school has been gradually trained to be compulsive and also to be concerned about his adequacy. That is to say, the personality scores may be a biproduct of overachievement in that setting. Despite the inadequacies of the design and reporting of the study, it raises a question that certainly should be studied in further work on reading.

A relatively recent study (Kight and Sassenrath, 1966) stimulated by the Grimes-Allinsmith work compared different types of college students in programmed instruction on the topic of measurement. This was not an experiment; there was only one treatment, an instructional booklet requiring about 6 hours to complete. Information was available on the Test Anxiety Questionnaire and a projected measure of need for achievement. Four groups of subjects were contrasted. Anxiety had little to do with time to complete the program, whereas there was a significant relation for motivation to achieve, those high on this variable completing the program much sooner (!). With regard to errors during the program, the High-High group did well, the Low-Low group did badly, and the other groups tied half-way between. On an achievement test given immediately after the program, the High-High group did best, the two groups low in motivation to achieve did badly, and the group High in motivation to achieve and Low in anxiety did moderately well. This does not support the initial hypothesis that programmed instruction, being highly structured, would work especially well for more anxious subjects. The more anxious subjects did have some advantage, but the strongest effects was clearly associated with motivation to achieve. We shall have more to say about that variable later.

One oversight in designing this experiment makes the weak results on anxiety highly equivocal. It has generally been found that test anxiety is strongly related to ability, with better students reporting less anxiety. This need not reflect any deep personality manifestation; the good student does not have to worry when he is facing a test. These investigators made no effort to examine the regression of outcome on ability, though they did equate their three groups on a pretest over a specific content of the program

If it is true that high test anxiety was associated with lower intellectual ability, then the modest differences favoring the high anxious would be intensified by partialling out ability. We may recall earlier remarks about curvilinearity; it is possible that the low-anxious group in this study was unmotivated to the point of apathy, or that the High-anxious were at a peak so high as to interfere with performance.

A fifth-grade study by Campeau (1965) used programmed instruction, giving one group feedback to assist in the correction of responses, and the other group no feedback. The number of subjects was small, especially since the analysis was performed within sexes. Only persons at the extreme of the anxiety distribution were used; the dependent variable was a post-test score with initial IQ partialled out. There is no reason to think this analysis is incorrect, but/^{there} would have been some advantage in treating both IQ and anxiety as predictor variables in forming regression planes. For girls there was a significant interaction, with those high on test anxiety doing distinctly better when given feedback and distinctly worse than the low-anxious when given no feedback. This was also found on a retention test. For boys, the relationships were not significant, and on the immediate test there was essentially no effect. Reporting is inadequate. It is uncertain that the Low boys are similar to the Low girls; it is often found that girls are considerably higher in test anxiety and it may be that a girl's Low score matches a boy's High.

The author's interpretation is that withholding feedback intensifies motivation by maintaining a certain incompleteness. That is to say, the "no feedback" situation is more challenging and more stressful. Alternatively, one could perhaps say that the provision of feedback provides greater structure, leaving the person much less on his own resources. The essentially negative result for boys is not explained.

A pair of studies done in England (Leith and Bossett, 1967) used measures of anxiety and extraversion. There were 60 10-year old subjects, divided into four groups on the basis of those variables and further divided among four treatments, so that there were 4 subjects per cell. The treatments were instructional programs affording varying degrees of guidance and discovery. The analysis of raw gains is entirely unsatisfactory. Since intraversion-extraversion seems to produce no effect, we may interpret the results in terms of anxiety alone. These results/^(Figure 15)are strong, but mystifying.

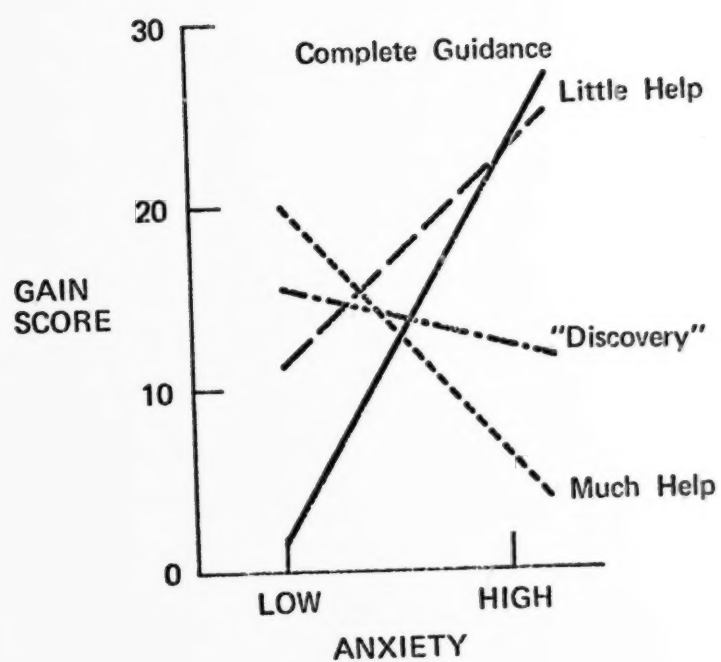


Figure 15. Results from Leith-Bassett study of 10-year olds.

The fact that we are dealing with gains scores in small groups is undoubtedly an important part of the difficulty. The discovery treatment provided little help. (Indeed, since it presented stimuli in a random sequence, it could scarcely be called a discovery treatment. It would be consistent with other results to find the high-anxious responding best with complete guidance and badly when given less assistance. We are inclined to take this study as supporting the usefulness of research along these lines, without accepting the results of this analysis.

The second study, with college students, used two instructional programs which differed in a great many respects, one of them was said to be a reception treatment and one a discovery treatment. While the low-anxious did better on this, there was no interaction with anxiety. We are told that there is a significant relation between extraversion and learning, with extroverts doing much better under guided discovery. One cannot be confident in assigning meaning to variables in these studies, since British questionnaires often include in introversion what Americans include in anxiety. In any event, the authors do not report in a satisfactory way, and one would be hesitant to say that a relation has been established.

Ryan, 1968, introduced structure in another way, by providing advance organizers. One group had advance organizers each day followed by programmed instruction, one had an organizer prior to any instruction and then programmed instruction each day. One group had both initial and daily organizers, and the fourth group had no organizer. No interactions were found, though the organizers were apparently helpful to learning. Susan Crockenberg, one of our associates, suggests that the failure to find the expected effect in this study may occur simply because programmed instruction provides sufficient structure to serve the psychological requirements of the more anxious subjects.*

Another type of structure was offered in a small study by Gifford and Marsten (1966) where 31 children were given pre-training in taking tests or were introduced to the task only by/ ^{being} given general directions. Performance was strongly related to anxiety when only simple directions were given.

* We are indebted to Mrs. Crockenberg for many of the references and notes in this section.

The anxious students took much longer. With test-taking training, the more anxious students took slightly longer, but the regression of reading time on aptitude was negligible. This should not be regarded as a learning study in the same sense as most of the others. This is more nearly a case of "tuning" the subject so that he is able to display the reading ability he already possesses.

It is interesting in this case that the pre-training had the effect of causing the low-anxious children to spend longer time in reading, and presumably to read more carefully, because of their now understanding the necessity for comprehension. Mrs. Crockenberg notes that the highly-anxious student is reputed to be preoccupied with his own irrelevant responses (Mandler and Sarason, 1952), and that pretraining or other structure may simply be serving to bring his attention fully to the task in hand.

The next study to be considered (Neale & Katahn, 1968) also involved tuning subjects for efficient test performance. Subjects were, or were not, allowed to determine the order in which tests were to be taken. Among those who had no control over the order, some were told the order in advance and some were not. It was assumed that uncertainty would lead to greater arousal, but in fact the performance (on a digit-symbol task) was strongly related to anxiety under the condition where subjects had full control, the less-anxious subjects doing better. There was a weak tendency for high-anxious students to do better when given no control. But Mrs. Crockenberg points out that the absence of power by definition provides a structured situation. The subject has no responsibility for decision-making; he may be under less stress and be more ready to concentrate on doing the task.

To summarize this series of abstracts, then, we can say that there are repeated hints that students who are above the class median in anxiety respond better to "more structures" treatment. But structure is defined in various ways in these studies, and the class median on anxiety floats along the scale. The studies almost invariably use extremely short instructional treatments, and there is a most haphazard mixture of defining variables for the personality construct: manifest anxiety, test anxiety, introversion, compulsiveness, etc. There is also a distressing variety of ways of taking mental ability into account in the designs for these studies, and this is particularly critical because of the usual correlation between reported anxiety and past success in school. On the whole, then, the concept of structure has served a number of investigators to lead them toward

a relation that repeatedly reaches nominal significance. We need considerably more systematic work, with careful analysis of what "structure" is to mean; with careful control of other variables; with adequate samples and adequate statistical analysis; and with treatments comparable to classroom instruction. The incomplete work by Stallings/^{and Snow} (section E) comes nearest to this kind of investigation, but what has been done so far must be regarded as pilot work. From the point of view of the present concern,

that work suffers from the decision to use a personality variable (learning avoidance) as an outcome, without obtaining pretest information on personality. Measurement of personality at the primary level is not easy, but some workable procedures are available.

Task difficulty as treatment variable. We turn now to another prominent line of research, on the postulate that anxiety interacts with task difficulty. We have already seen one or two studies in our earlier examination of literature on programmed instruction where difficulty and anxiety were simultaneously under consideration. The majority of the studies of the problem have used laboratory learning rather than instructional material. The original impetus for the work came from the theoretical argument that drive acts multiplicatively with habit strength to determine what response will be made. It was proposed that anxiety as reported in a questionnaire could be taken as a measure of typical drive level. If only one response is called for, then high drive should produce strong and effective response. If there are a number of competing responses, that might be disadvantageous, arousing incorrect responses as well as correct responses. The most direct evidence came from studies of eyelid conditioning and simple word-association, where anxious subjects did better on difficult paired-associate learning and complex mazes. Particularly valuable in confirming this theory is the Katahn-Lyda (1966) word-association study, in which the high anxious did best when the keyed response was high in his personal repertoire of responses, not otherwise. Lows were not affected by the change in response availability.

McCandless and others, 1956, required the subject to learn a series of button-pushing responses that would turn on lights. There were easy and difficult combinations. On the Children's Manifest Anxiety Scale, those scoring 18 and above did relatively better than others on the easy version, and less well on the difficult version.

Covington (1967) used a problem-solving task as his dependent variable, and gave attention separately to the sheer output of possible relevant ideas, to the quality of ideas, and to the quality of the idea the subject judged to be best. General Anxiety seemed to have much stronger correlations with the difficult task of judging the quality of ideas than with sheer fluency of production. Correlations for Test Anxiety were smaller, and only the correlation with total unevaluated ideas was significant. The more anxious children produced poorer ideas and were less able to judge quality; the test-anxious children apparently produced more ideas. This tends to support the idea that a demand for quality fits the aptitude of the low-anxious subject, but this is not an ATI study with experimental manipulation.

In two studies of college students with difficult and easy serial learning, Katahn (1966), got results that cause difficulty for the hypotheses of other investigators. In one study there was no relation of outcome to anxiety, and in the other study the high-anxiety students did best despite the difficulty of the task. In this study there was no easy task, and it is possible that Katahn got an unexpected result simply because his students did not find that task as difficult as he expected.

Smock (1958) studied children with extreme scores by an anxiety scale. Information processing in the two groups appeared to be different, though no evidence directly bearing on learning was collected. The more anxious subjects responded quickly on "Gestalt closure" problems, and also tended to respond in other tasks before they had sufficient information. This is strong suggestion that when faced with a somewhat confusing situation, the high-anxious subject confuses himself by generating premature hypotheses.

The difficulty of tasks was crossed with the further variable of massed and distributed practice on two kinds of intellectual tasks by Korchin and Levine, 1957. Anxious subjects were helped considerably by distribution of practice, whereas this was unimportant for the low-anxious. Within the range of practice schedules used, the Lows were superior throughout the range. High subjects were poor even on the easy task. Here again, information processing seems to be at issue.

Eysenck (1967) discusses the possibility that extroverts, whom he sees as less aroused (like Spence's low-anxious), will be superior in short-term recall tasks (perhaps up to one hour) and inferior on long-term recall of paired-associates. Introverts (high-anxious in American terms) have, he suggests, the opposite pattern. This seems moderately consistent with Korchin and Levine, since spaced practice should show the long-rememberers to advantage.

Eysenck goes on to discuss work (see Eysenck & McLaughlin, 1967) in which four groups are assorted according to their postulated drive level that accompanies their neuroticism and extroversion score. The chart he gives is spurious, drive levels being adjusted to make good "arch" curves. But if we accept his order and not his scaling, the trials to criterion in paired-associates learning are:

	Stable extroverts	Neurotic extroverts	Stable introverts	Neurotic introverts
	Low drive		High drive	
Easy List	83	27	54	85
Difficult list	82	139	148	167

Thus the added stress of the difficult list puts the low-drive stable extroverts at the peak of the arch, and the lower stress of the easy list favors the neurotic extrovert. The implication is that a still less-challenging task would favor subjects still higher in drive. The reader's attention is directed to Eysenck's many other remarks, admittedly speculative, on the possible nonlinearity of ability-personality relations and the need for studies of learning with various parameters under systematic control.

Reinforcement as a treatment variable.

A further line of research has employed a style of reinforcement as a treatment variable.

Subjects at the extremes of the test anxiety distribution worked on the John-Rimaldi reasoning task under three conditions: evaluative, neutral, and play (Blatt, 1963). Low anxiety students performed better on the task, but it is conceivable that this reflects only their superior reasoning ability. This study would have been improved by a design that permitted a regression analysis with both variables simultaneously, but the extreme-groups design allowed only 10 subjects for each of the six cells of the design. When the effects for such extreme cases are slight, the results are not impressive. Moreover, as familiarization takes place the inefficiency of the anxious student disappears. The neutral, nonevaluative condition was rather unsuitable for the high-anxiety subjects. The evaluative condition where an emphasis was placed on the quality of the student's performance seemed to get better results from the high-anxiety person. The effects are untrustworthy, because of the small number of cases and the compounding of anxiety with ability. They do tend to reinforce some of our earlier remarks about the importance of continuing an experiment past the familiarization phase.

Rin (1965) tested twelve-year-old subjects/^{on} suitable versions of the Eysenck inventory, neuroticism and extraversion being measured. (The neurotic score has a substantial correlation with the anxiety score of American questionnaires.) Six trials of a cancellation task were administered under conditions of praise, blame, or no comment, for a series of six trials. It appeared that neither praise or criticism had a constructive influence on performance of the nonanxious children. With high-extraversion criticism had a particularly valuable effect. The data seemed to suggest that criticism got superior results with all kinds of subjects. Experiments of this sort are of little value, however. A similar study by Konstadt and Forman (1965) classified 38 children on the basis of the Embedded Figures Test, which can be seen, as one chooses, as a measure of dependency or as a measure of fluid ability. There were two examiners, one using an approving manner and one a disapproving manner; each examiner was first for some children. The approval condition got good results from

the dull-dependent children (work output on cancellation being the criterion), and the disapproval condition got bad results, regardless of order. The approval condition got good results from Highs only if it came first; results were very bad when it came second.

Experiments of this sort have little value, because of the brevity and novelty of the treatment. In a classroom one would expect the students soon to be habituated. Moreover, it is hard to believe that any teacher would administer a steady diet of either praise or blame to an individual. Basically the same criticism can be made of a study of attitude change by Greenbaum (1966), but the study was conducted with considerable ingenuity and may be worthy of examination. Students were directed to make a speech on civil defense in a direction contrary to their opinion. They were criticized after the speech: the criticism could be strongly favorable, strongly unfavorable, or intermediate. A complete analysis was given for students classified with respect to "need for approval" (defensiveness). There was marked attitude change for the persons low in the test of defensiveness, regardless of whether the comments about the speech were positive or negative. There was no such relation for self-esteem. The results for a second group of subjects are extremely complex and, with an extremely small sample in each cell, one can have little confidence that the inconsistencies among results are interpretable. In this series subjects were ostensibly allowed a choice of the attitude to be expressed in their speech, but in fact they were subjected to strong social pressure to present the side of the story they did not believe in.

A study that at first glance appears to be a study of praise or blame (Iwahara & Tanabe, 1963) is actually more nearly a study of how subjects use reinforcement information. In a simple learning task, the subjects were reinforced with the words "right" and/or "wrong." One group of subjects (RW) was told whether a response was correct or incorrect. A second group (W) was told "wrong" after every incorrect response, and nothing was said after a correct response. A third group (R) was told "right" after every correct response. Thus after every trial the subject knew whether he had made a correct or incorrect response. There is a strong disordinal interaction. For groups RW and W the means were 0.60 for

anxious students and 0.46 for the low-anxious; for group R, the means were 0.45 and 0.46 and 0.67 respectively. This score was the number of responses to reach criterion, and therefore the R treatment was well-suited to the anxious subjects, and the RW treatment to the non-anxious. Now in this measure, the W treatment was also superior for non-anxious subjects, but the difference was somewhat weaker when a number of errors was counted. On the whole, however, one would conclude that the low Anxious made good use of signals of wrongness, and the high anxious were impeded by these signals. Reinforcement, it should be remembered, was conducted in a neutral manner, so that the statement "wrong" was not to be considered a personal criticism.

With this finding in view, it would be interesting to return to the old "negative practice experiment" of Knight Dunlap⁽¹⁹²⁸⁾. It will be recalled that he found persons who deliberately practiced an incorrect response able in the end to give correct responses perhaps more efficiently than when they practiced directly. One might infer from the Iwahara-Tanabe results that negative practice is particularly useful for low-anxious students. Again, one would want to ask about the cumulative effect of any particular kind of learning. Few experiments have carried a novel treatment long enough for one to be sure that we are dealing with more than an initial adaptation to the novelty.

Constructive motivation

In contrast to the broad pattern of tension and fearfulness around which the preceding section was organized is a pattern of ambition, energy, and goal-seeking that we can assign the blanket term "constructive motivation." In terms of the Semantic Differential dimensions, this is a good-strong-active pattern. It includes the concept of need for achievement; the reader will recall that Atkinson contrasts anxiety and need for achievement, the one reflecting fear of failure and the other hope for success. We avoid any attempt at refined delineation of variables in this area, as the results are confusing even with regard to a single measure, and there is no basis for identifying the range of traits that enter into similar interaction patterns.

The most active attempt to identify interactions of constructive motivation is that of McKeachie, who with his associates has made comparisons of different styles of college instruction, as they interact with student personality. This research has not been given a full technical report. In 1958 McKeachie described early results in a speech, and a long lecture given to the Nebraska Symposium on Motivation (1961) recounts more extensive studies. The only subsequent report (McKeachie et al., 1966) deals primarily with need for affiliation.

Brief reference, without tangible data, is made to a finding on differences in "feedback" in instruction. Where the instructor was explicitly evaluative, announced tests in advance, and the like, performance of students high in n ach and low in anxiety declined. Apparently Low-low students did well with high feedback. (There is no comment on the High-high group). McKeachie's total report, involving other motives, leads him to a strong conclusion that ATI have been demonstrated. The inconsistency of results and their generally small magnitude should be kept in mind, however. Nor can one take findings with full seriousness in the absence of a solid technical report that indicates, among other things, the total number of relationships from among which the "significant" results came. (Such data may appear in USOE contract reports we have not seen. These contract numbers are SAE-8541 and 4190-001.)

Some instructional procedures give the learner more freedom and more responsibility for planning. Patton (1955; see McKeachie, 1958) found that when a college psychology class was organized in this way the students who took most responsibility learned the most and had the most favorable attitude. And such students tended to have high n ach and little dependence on authority.

One table in the 1961 McKeachie report also involves n ach. 278 students out of 583 earned grades of A or B in psychology; but in the cell where n ach was high (aptitude) and the number of achievement cues presented by the instructor was low (treatment) the proportion was 82 out of 138. This interaction is significant. The differences are said to be stronger for men. The superior grades are explained well enough by the argument that when the instructor does not induce motivation, the Highs

motivated themselves. What is hard to explain is why the Highs did no better than the Lows when motivational cues were provided by the instructor.

A rather remarkable technique was employed by Domino (1968) to identify treatment variables existing in college instruction. Domino was studying the success of college juniors, and interviewed the instructor of every course one of his subjects was taking to identify the presence of such features as emphasis on material to be memorized, keeping of attendance records, objective examinations, etc.; these more constraining characteristics were thought to identify reward and require conforming behavior. A contrasting group of characteristics that reward independent behavior on the part of the students was also probed for. Some 73 courses were classified as "conforming" and 32 as "independent." Grades the student had received were separated according to whether they had been earned in the former or latter kind of course, so that he had two grade averages. In the end, there were four groups of 22 subjects each, matched for sex and score on a nonverbal mental test. The four groups were defined as coming from the extremes of the distribution on two California Personality Inventory variables: Ac (achievement through conformity) and Ai (achievement through independence). One would question the careful matching on sex and mental ability, since these are very likely correlated with Ac and Ai; the two variables could better have been retained as covariates.

The Domino data produced grades averages as follows (the average in conforming settings being given in the first row of each cell, and that in independent settings in the second):

	Low Ai	High Ai
High Ac	2.7	3.0
	2.4	3.3
Low Ac	2.3	2.5
	2.1	2.7

If there is an interaction, it is in the tendency of High Ai students to do better in situations that reward independence, and for Low Ai students to do better in conformity situations. There is no interaction involving Ac, though it appears to have some validity as a general predictor. Even though

Domino stated that he was interested in "differential achievement", his significance tests do not examine differences between kinds of courses, but rather differences between subgroups on the same courses. The standard error of any one of the means given above is in the range 0.1-0.2; but we have no information about the correlation of the two grade averages for the same person, which might reduce the error term in a test for interaction considerably. Even if the interaction is significant, its magnitude must be regarded as small when we recognize that Domino is contrasting students from the far extremes of the Ac and Ai ranges. It is most regrettable that, having gone to the trouble of classifying courses, Domino did not use test scores for all 348 of his original subjects and apply a regression analysis with sex, ability, Ac and Ai as predictors. The weakness of his relations is not a reason for abandoning his approach, since his working hypotheses about relevant variables were crude, his instructors may have been biased as informants, and the CPI scales have only moderate construct validity. On the whole, the finding encourages more powerful studies in the same vein.

A great volume of work, much of which is summarized by Atkinson and Feather (1966) and Heckhausen (1967), has shown interactions of n ach with treatment conditions in experiments. These have usually been studies of risk taking and not instruction. A considerable theory has emerged that might be the basis for formulating hypotheses about educational treatments. The one study that requires our attention examines such an hypothesis in the classroom. O'Connor, Atkinson, and Horner (1966) inferred that homogeneous grouping would serve pupils high in n ach and low in test anxiety. Various sixth-grade classes were used, some of which had had homogeneous grouping the previous year. It was concluded that students relatively high in n ach showed greater interest and enhanced learning when grouped by ability. The anticipated decrement in high-anxious students was not found.

The details of the analysis are hard to follow. A motivation score was formed by subtracting standard score on anxiety from standard score on n ach. The first table, using posttest achievement scores, must be interpreted as showing a definite absence of interaction. Those higher in the motivation score did better, and the ungrouped classes did better. But the line relating outcome to n ach for the ungrouped classes is above that for grouped classes at all motivation levels. The interaction reported as significant is based on raw gain scores (year-to-year gain in achievement). This technique is simply not acceptable. Even if the results are taken at face value, the interaction is disordinal, and this time the gain

score indicates that the grouped classes do better. This occurs partly because IQ is treated differently in the second analysis. It would be correct to consider four pretest variables side by side: n ach, anxiety, mental age, and achievement pretest; then an analysis of the posttest data would mean something.

As for the interest variable, there does seem to be a disordinal interaction. In homogeneous classes interest is sharply related to n ach; in heterogeneous classes, there is no relation. This analysis too would be improved by bringing additional pretests. The meaning of the finding (assuming that it can be replicated) would depend on how the teacher handled grouped classes -- e.g., to what extent competition was stressed.

A small study by Ryan and Lakie (1965) may be given brief attention. Their question was whether competitive conditions would benefit some learners more than others. The task used was placing rings on a peg, guided by a mirror. Subjects first worked under neutral and then under competitive conditions. Ss were classified on n ach and on manifest anxiety. The number of cases per cell was ridiculously small (about 7) and the reporting poor. We infer that the cell means were about as follows (initial condition mean under noncompetitive condition given first):

	High anxious	Low anxious
High <u>n</u> Ach	23;32	22;34
Low <u>n</u> Ach	23;29	23;32

While the gains are evident, one cannot take seriously a significance test on the raw gains. Perhaps competition is especially beneficial in the upper-right cell where motivation is most constructive, but one could not defend the conclusion on the basis of this research.

Other motives

A few studies in the McKeachie series deal with affiliation and power motives. The latest paper, summarizing prior work, ends with the conclusion that in three separate studies men high in n affiliation made relatively better grades in classes where there were many affiliation cues, and Lows did worse. Results for women were not consistent. Subjects were selected mysteriously (perhaps from extremes of the distribution) and classified differently in the different studies. We may consolidate results by pooling "Highs" and "Middles" and indicating the percentage of mean earning A or B in the course. For Highs, the percentages were 54 in high-affiliation

settings and 46 in low affiliation settings; the percentages were reversed for lows: 43 and 55. This effect is made more dramatic than it actually is by the selection of cases. It is to be noted also that in two of the three studies an objective test of achievement did not show a similar interaction and in the third study an essay test gave no significant interaction. In sum, the effect may tell us more about instructors' grading practices than about student performance.

For n Power we must turn to McKeachie (1961), where we find that men high in power motive tend to get their best grades in classes where volunteering is encouraged. This finding did not emerge for women. Again, the information may have to do with grading practice and not learning.

A questionnaire measure of "dependence proneness" has been used in the research of Flanders (1965). It was supposed that students high on this variable would profit more from teaching that was "direct" (more lecturing, more use of authority in class control) to "indirect" (more encouragement, acceptance of student feeling, more use of student ideas). Junior high-school classes in social studies and mathematics were selected from the high and low extremes of a distribution on a scale where students reported their teacher's style. There were achievement pretests and posttests, and an IQ for each subject. The author finds no interaction with pretest, IQ, or dependence as predictor. (While the report does not mention a test on regression slopes for the predictor, the fact that analysis of covariance was applied to test the main effect implies this.) The author claims a significant main effect favoring classes taught indirectly, but the wrong error term appears to have been used. A mean difference of 2-3 points between teaching styles is not significant when the within-class s.d. is 6-7 points and there are 7-8 classes per treatment.

One might make some proposals to refine the analysis, but it appears unlikely that these would alter the essentially negative finding on interaction. Flanders chose to residualize outcome scores on the basis of the pretest only. Formally, use of estimated true scores on the predictors is called for, since the aptitude means (at least of IQ) of the indirect and direct classes differed; but a difference of about 0.3 s.d. ought not to affect results greatly. Second, we would recommend treating all predictors simultaneously in an overall test for interaction.

Snow, Tiffin, and Seibert (1965) used a variety of personality and attitude measures, along with ability and prior experience variables, to investigate interactions with learning from filmed vs. live lecture demonstrations in college physics. 437 freshman engineers participated. The alternative treatments represented the major portion of a semester's work for one course. Again, the full power of the data was not capitalized upon; blocking procedures were used to form a series of three-way unweighted means analyses, where regression methods would have been more powerful and more parsimonious. Results using immediate recall criteria indicated that two personality variables, Ascendancy and Responsibility as measured by scales of the Gordon Personal Profile, interacted with film vs. live treatments. The more ascendant, assertive students and those appearing relatively irresponsible, profited more from live demonstrations. The more submissive and more responsible students seemed somewhat better off with film. Whether these findings relate to earlier notes concerning constructive motivation or to work with motives such as affiliation or independence cannot be judged at present. Personality variables in all this work remain inadequately measured and the findings stand unreplicated. Additional complex interactions were obtained using prior knowledge of physics in combination with attitude toward instructional films, and verbal and numerical abilities, but these were of doubtful use. Of more importance, potentially, was the finding that prior experience with the instructional method (i.e., film) might provide a basis for ATI. The suggestion follows that learning-to-learn effects may be represented in such self-report experience variables.

Another study, by Tallmadge, Shearer, and Greenberg (1968), included the Gordon Personal Profile and the Kuder Preference Record, along with several ETS Kit tests, in comparisons of "inductive" and "deductive" instruction. Two one-day training courses (Transportation Technique and Aircraft Recognition) were offered using each of two methods. One method used an example-rule form, where examples, questions, and partial information about rules for problem-solving were given by the instructor. The other provided straight exposition of rules and their application to problems. Subjects were 231 Navy enlisted men. The results involved three-way interactions including both method and content as well as complex combination of aptitude scales. While continued interest in the Gordon scales and in some

Kuder interest patterns (e.g., scientific vs. musical) may be justified on the basis of this study, the complexity of the results and the unavailability of regression data for single aptitude variables makes detailed interpretation impossible at this time. Two general comments made elsewhere in our report can be underlined again here, however. First, detailed process analyses will be required to show similarities between treatments that are nominally different, and vice versa; method X content interactions may signal inadequate task description, as Tallmudge et al. note. Second, narrowly differentiated cognitive factor tests again prove of little value in ATI work.

A particularly miscellaneous study by Cahoon et al. (1968) found an apparent interaction in an unlikely place. Students were chosen who had high Kuder Mechanical interests and low Kuder Literary interests, or vice versa. A 19-frame instructional program was presented, either in the form of printed text or with the aid of a teaching machine. Students in the High Mechanical/Low Literary subgroup were benefited by the machine, and vice versa. Since the instruction was trivial in duration and in content, perhaps it is best not replicated.

A study of teacher differences

The one really impressive study in the personality area is that of Heil et al. (1960), where fifty elementary teachers and their classes were studied. Questionnaire data led to a typing of teachers as relatively spontaneous, or orderly, or fearful in their management of classes. Pupils were categorized, as strivers, or docile conformers, or opposers. Average achievement for pupils of each type under each type of teacher was calculated; a correction for class differences in IQ was made. It was necessary, in order to reach consistent results, to divide teachers in each style group into superior and inferior subgroups; the superior teachers were observed to be more warm and democratic. The results are complex, and can be no more than tentative when one is reduced to six teachers in some cells. The information can be put in the form of mean adjusted achievement, with these findings:

The striver type of pupil did well over all teachers, and did:

- a) very well with the spontaneous superior teacher
or the orderly teacher
- b) only mediocre work with fearful, inferior teachers.

The docile conformers did about as well as the strivers over all teachers and did:

- a) extremely well with spontaneous superior
or orderly superior teachers
- b) mediocre work with fearful teachers
- c) distinctly bad work with spontaneous, inferior teachers.

The opposers did bad work, all teachers considered, but did:

- a) very well under orderly teachers
- b) extremely badly under spontaneous teachers.

The results scarcely allow one to raise questions of statistical significance, when teachers are the unit of sampling and some of the classification is post hoc. Nonetheless, this is a most provocative finding. It argues strongly that research intended to identify "what kind of person makes the best teacher" is futile. It suggests some rules for pupil assignment. The fourteen orderly teachers got better results, all types of pupil considered; the fearful teachers got rather poor results. The spontaneous teachers (the smallest group) were about equal on the average to the fearful teachers but their variation over kinds of pupils was spectacularly large. The resistant pupil is evidently cut adrift in the spontaneous classroom, and goes onto the rocks. The dutiful pupils are evidently swept to unusual peaks of achievement, where the spontaneity is accompanied by warmth. It is hard to visualize a cold, undemocratic, spontaneous teacher; the few teachers of this kind were clearly bad. They did least harm to the striving pupil. If the Heil data were to be confirmed, the proper school policy would be to capitalize on both main effects and interactions. Teachers who are warm-democratic-responsive and either orderly or spontaneous in style should be sought or developed. Those with the orderly style should get all the "opposers"; they can handle the other kinds of pupils well but the spontaneous teachers get even better results from them.

On the whole, studies in the personality area are disappointing and unencouraging. The studies have often been flimsy; a small number of cases, tested by a single instrument of uncertain interpretation, given a short and artificial treatment and measured on variables not too significant from an

educator's point of view. The studies carried out in classrooms have not been controlled (with regard to nature of treatment and student assignment), and the effects have generally not been strong. There are good reasons for the character of these studies. They have often been part of general programs of work on issues in the personality area rather than of work on instruction or even on the psychology of learning. At best, it is difficult to control treatments. Moreover, the state of thinking about personality variables as they relate to instruction is in a primitive state, so that planned treatments are probably premature.

Of the findings we have examined, the only one that is supported from many sides is that making instruction more difficult or more demanding may interact complexly with anxiety, so that for each level of anxiety there is a best level of instructional pressure. This notion has not been carried into genuine classroom experimentation extended over a reasonable period of time, nor has there been a sufficiently direct demonstration of the supposed curvilinearity of regressions in learning situations. All too often, the argument derives from results that fall on one or the other side of the "arch", so that the second leg is an entirely hypothetical construction.

H. Individual Differences in Instruction:

Future Prospects

Ways to adapt instruction

There are a number of ways in which educators can cope with the fact of individual differences among pupils. They range from procrustean methods that involve little adaptation, through intuitive and little tested rules for adaptation, up to, in principle, tested rules derived from theory.

The least responsible solution is to fix the curriculum and method of instruction, and to "adjust" through initial selection and through allowing for dropouts. This eliminates the worst absurdities of trying to force instruction down the throat of a pupil who cannot or will not respond to the instruction, but it is not a constructive solution.

The educator of the past generation has been willing to adapt instruction to individual differences. We may distinguish between two broad kinds of adaptation. One is to choose different educational goals for different persons, and the other is to choose different educational means toward the same goals. The former no doubt is valuable, particularly with respect to developing the person's capacity for self-expression in work and leisure. It cannot be the only policy, however. There are some educational goals that all pupils must move toward, and attain to the greatest degree that educators' ingenuity permits. The easy escape of shunting some pupils into a "nonacademic" curriculum cannot be tolerated, so long as proficiencies formerly considered "academic" are necessary for success and social contribution. This is made clear in the black protests against grouping systems that hold some young children in a simplified and unenlightening program, and against the selectivity of colleges that make provision only for applicants likely to succeed. The demand is that educators invent new programs to open opportunity to persons who would not succeed in attaining traditional goals in traditional ways.

Our concern, then, is with adaptations in method that will fit instruction to the relevant characteristics individuals bring to the classroom. Teachers in recent years have done this by a variety of tactics: diverse reading materials to suit children with different

skills and interests, diversified projects calculated to appeal to different pupils, individualized remedial work, and so on. The past decade saw the advent of programmed instruction, offering new possibilities. Linear programmed instruction offered the same instruction to all, but allowed for differences in rate of completion. Branched programs, and later computer programs, offer the possibility of mechanizing adaptations of the sort the skilled teacher introduces more haphazardly. For example, remedial loops to fill in subskills are now a well-standardized procedure. Adjustment of the length of time a pupil spends on a certain kind of drill is also automatic, being determined by systematic monitoring of his progress and application of a decision rule. (One rule, when progress is too far from the normal range, is to summon a teacher who can apply judgment going beyond the programmed rules.) A variant of the same strategy is found in individually prescribed instruction, Project PLAN, and other schemes reminiscent of the Dalton and Winnetka plans of a generation ago. What the pupil should work on is determined in collaboration by the pupil, his teacher, and the system of rules; an important part in the decisions is played by extensive measures of his achievement.

All these procedures are essentially atheoretic. The basic concept is that each unit of instruction lays down a baseline of proficiency on which the next unit can build. Conversely, one can specify the proficiencies needed to master a new unit, and then, after taking inventory of what the pupil can do, can put him through remedial work. This concept of mapping hierarchies through the subject matter has undoubted value. It is best suited, however, to achieve training in well-specified subject matter. It clearly can be made to work to teach skill in manipulating decimals. It is much less applicable to the broad concomitant outcomes of instruction: development of mathematical intuition, comprehension of mathematics as a system of thought, and the like. Teachers can make rough guesses as to the sorts of activities that will best promote mathematical thinking for a particular pupil, using general concepts about ability and motivation that they have distilled from past experience. System designers can collect such wisdom and to some degree formalize it and supply materials to help pupils of various types. But these plans have a much less evident

rationale than plans to check mastery of whole-number computations before bringing in decimals.

Systems of the sort just discussed are limited because they are designed on the basis of premises that cannot be validated in the setting of the complex system. To be sure, the system as a whole, like any intact curriculum, can be evaluated. But a plan that works fairly well may still embody assumptions that would not hold up under close scrutiny. The function of research, theory, and systematic development is to expose mistaken assumptions and provide more finely designed tools to do better what the intuitive methods do only reasonably well. Evaluation of a particular course of instruction can do little to formulate and test general policies for the design of other instruction. We can see no short-term solution to the problem of individual differences save artistic design of alternative instructional schemes.

The long-range requirement is for an understanding of the factors that cause a pupil to respond to one instructional plan rather than another. These plans should differ in more than the amount of time devoted to specific drills. The range of instructional procedures open to the educator is enormous -- individual projects, workbooks, teacher-monitored problem-solving, group projects, discussion, etc., etc. The development of new media greatly extends the range of methods and also extends the capability of the school to administer flexible and diversified programs. There is no reason to assume that an eclectic mixture of all methods will serve all kinds of pupils, or that the choice of methods is a function of subject matter alone. On the contrary, there must be some kinds of pupils who respond best to group discussions, and others who do much better by themselves. The same is to be said of all the parameters of instruction: level of reading and other kinds of comprehension required by the presentation, sternness or permissiveness of supervision, degree to which competition is introduced, etc.

The most conscious planning to match students to diverse instruction is found in the attempt of colleges to find distinct styles; sometimes this is achieved by organizing independent colleges within the same university. To this point, the only tactic for achieving a

fit has been self-selection. It is assumed that if clear enough information about the college is made available to potential applicants, the students will recognize if they are suited to it. One can have no confidence in such self-matching. The high dropout rate of an atypical college like Reed, for example, reflects the fact that many students who think they will like a novel program find it raising anxieties they cannot tolerate. Likewise, many a traditional program in the sciences has lost students who thought they would find adventure and found only tedium. All experience in guidance leads us to recognize how little insight the student has, even at the college level, into his own motives and capabilities. At lower levels it is even less likely that allowing the student a range of choices will serve. Hence, as differentiated programs are introduced, there is a clear need for systematic investigation as to the kinds of students who thrive in each -- that is, for the study of ATI.

Before leaving this general discussion, we record one warning. The ATI conception may do disservice to education if it ultimately programs students only into work they can handle competently. It may be perfectly true that a student of Type Q gains relatively little when taught by a discussion method. This certainly argues against a school where he must learn his history and his psychology through attendance at discussions. But if his failure to profit from attendance at discussions is due to shyness, or disrespect for the opinions of others, or an inability to process information received in a somewhat disorganized auditory form, it would be wrong to let him carry those deficiencies throughout life. Any significant interaction implies that the person has greater aptitude for one treatment and less aptitude for another. However well we cope with the short-term instructional problem by employing the former treatment, we should not fail to consider how the inferiority may be removed. An inferiority that limits the student's response to a certain kind of instruction is also likely to limit him in social or work situations. To arrange the college experience of our Type Q student so that he has minimal occasion for group discussion would be to atrophy what limited skills he does possess. The problem for the educator is to find a way to encourage growth in ability to profit from discussions, without thereby sacrificing his

learning of subject matter that could be presented in ways other than discussion. Like all educational choices, considerations of time and cost will determine just how large a fraction of the student's time will be used for each purpose.

Requirements in ATI research

The reader who has followed us through the maze of our survey will have sensed that the ATI problem is a frustrating one. Most research of this kind in the past has been inconclusive, either because questions were badly put or because investigations have contradicted each other. Few or no ATI effects have been solidly demonstrated. Mapping out a theory can scarcely begin until a reasonable number of relationships can be asserted with confidence, to give the theorist a place to stand. And yet, in the absence of theory, empirical research degenerates into random trial and error. While trial and error, in sufficient amount, can be instructive, any one proper study on ATI is likely to be an expensive undertaking, and sheer empiricism is intolerably wasteful.

Many aspects of past research have contributed to its impotence. Plans have been laid on the basis of inadequate thought; experiments have been unduly brief and have carried no provision for choosing among viable counter-hypotheses, analysis has been weak and often incorrect, replication has been next to nonexistent. We are not alone in deploring the effort wasted in educational research by superficial and incorrect analysis of expensively collected data. Stanley (1967) demonstrates that the concepts and designs framing a study are typically more elaborate than the analysis, and the latter is inadequate. There is basic agreement between his proposals and ours, though we are more interested in descriptive summaries of data, construct interpretations, and validity generalization where he is more interested in application of formal statistical models. These faults are a mark of a field where investigators are just learning to crawl. The studies conducted within this project are surely not prototypes for future work; on the contrary, it was out of that experience that we arrived at many of our ideas about changes that should be made in future studies. We hope that investigators whose studies were used in the

preceding pages to illustrate faults will not be too distressed by our Monday-morning quarterbacking. Had we undertaken the same studies at the same point in time our work would have had equally grave faults.

We will allow ourselves one adverse comment about our colleagues, or rather about the times in which we live. Research is a hazardous undertaking, and any one line of inquiry is likely to turn up a mixture of the dependable and the undependable. It is through the sifting process of the community of scholars, contrasting studies with each other and bringing methodological and theoretical perspectives to bear, that the more dependable findings are identified and become ready for public use. This process is all too frequently bypassed by direct and large-scale advertising of conclusions from a limited line of inquiry. To publish results in book form is desirable, if that gives scope for full exposure of details of an intricate inquiry; but if the book is marketed commercially and accompanied by news releases and speeches dramatizing its value -- as has occurred for more than one of the investigations we have dissected -- a disservice is done. Even if the community does ultimately sort out the dependable contribution of the work, this tempered truth is unlikely to reach those who heard the initial intemperate trumpet blast. Indeed, by the time the sober version is out, a large number of uncritical investigators have adopted the overblown findings as a guide to their own research, and thus effort after effort is compromised. We see this tendency to premature dissemination as a sign of the pressure and-reward system in which educational and psychological research now operate, rather than as a sign that the investigators in question are less competent or conscientious than others. There are just too many pressures to do work quickly rather than as well as possible. The present report is an egregious example; left to exercise our own judgment, we would have spent a further year in refining it before release.

The fact that investigators take risks in embarking on kinds of research whose requirements are little known is what makes it possible to discover those requirements. They can be seen with some clarity in a synoptic undertaking such as ours, but we would not claim much for the degree of clarity we have reached. Our time and resources

have been much too limited to produce a definitive guide to future research. We hope that the serious worker will read through our comments in detail; he will need to evaluate them and formulate a preferred strategy for himself. The summary that follows here is incomplete, and no doubt presents the scene from only one perspective.

Treatments. The treatments used in past experiments have generally suffered from brevity and artificiality. The question before us is how students respond to instructional treatments. We are not going to learn this from studies that mimic laboratory experiments by presenting a single brief lesson repetitively until it is mastered, by confining instruction to a drill-and-practice mode with no explanation, or by introducing utterly artificial motivational procedures such as "blaming" the student day after day. We will need to collect data from instructional procedures that realistically progress through a body of material. The procedures should be good instruction, insofar as one can judge a priori or by tryouts. The instruction should be continued long enough that we know how the pupil learns after he is thoroughly familiar with the style of instruction; educational policy cannot be based on what the pupil does in his first encounter with an instructional style.

We foresee two rather different lines of attack. One is to use ongoing, distinctive educational programs. The McKeachie studies of college classes taught in different ways moves in this direction, as does the Herron study of chemistry classes. These studies will generally not be able to contrast randomly assigned groups, but they can employ large samples and collect data over long time periods. The work will probably be more informative if relatively intensive. A study in three schools is likely to produce richer information than a study in 30, just because a mail-order investigation cannot learn much about processes. One would want to know how pupils of different types respond to the instruction, as the course proceeds; this may well require observation as well as posttests. Any evaluation study has the opportunity and, many authorities would say -- the obligation to identify the kinds of pupils who do well and badly. If the evaluation study does not set up control groups (and it often should not) it still can stockpile findings that ultimately will bear on ATI theory.

The second kind of research will be experimental in the stricter sense. Special programs will be designed to contrast two or more treatments. There are serious constraints on such experimentation. Schools are unwilling to give up much time to experiments that are not contributing to the regular instructional goals. One cannot (as in the Stallings and Snow study) hold a pupil for long in an instructional treatment where he is not progressing as well as expected of him. One cannot afford to mount curriculum development for the purposes of an experiment. Moreover, the important requirement of reproducibility means that the programs need to be specified in much detail; programmed instruction is ideal in this respect, but costly to produce. We visualize, then, that the ideal treatment-set for this kind of experimental research is likely to consist in adaptations of some regular instructional material. Contrasting versions of instructional material or alternative programs of activities can be prepared without the investigator's undertaking to a major project of curriculum development. The school program is much less disarranged by the experimental requirement. Experimental treatments running for periods of a month or more may be feasible under this scheme, as is necessary if pupils are to learn to take advantage of a new kind of instruction. One particularly distressing feature of studies reviewed is that -- with a few rare exceptions -- each investigator has considered it necessary to write his own instructional materials. This is wasteful; more studies in which the same basic materials are used, perhaps with ad hoc modifications, would produce more solid evidence.

If we are not to have sheer empiricism, the choice of treatment variables will need to be deliberate and judicious. Such elementary hypotheses as that dull pupils should have small-step programs and spatially-bright pupils should be given much spatial material are evidently not worth much. The best hope for rational choice of treatments, at least in the near future, seems to lie in the direction of process analysis. Instruction is a process in which the pupil carries out a large number of actions, most of them covert. He attends, takes in, processes, and applies instructional stimulation; receives feedback (which goes through the same sequence of processing, more or less); and forms general attitudes and strategies affecting his future response

to this instruction and all instruction. Almost no writing on individual differences has attempted to identify or even to speculate about these processes. This was not possible a decade ago, when there was no psychology of complex information processing. But the stirrings in cognitive psychology (including social psychology, and differential psychology that goes beyond factor analysis) suggest that the instructional psychologist can profitably begin to think in these terms. We have found very little that can be called an illustration of this method. On a small scale, the Burton-Goldbeck (1962) study illustrates brilliantly what we have in mind; the Koran study is also pertinent, though the theorizing is post hoc.

Obviously, if we have working hypotheses as to the processes the student can and should use in a particular kind of instruction, we will coach the student to use those methods. We should not be interested in the effectiveness of the pupil's naive strategies. It follows that the ATI problem becomes one of identifying aptitudes that make it possible for some pupils to adopt a desirable strategy. Given contrasting treatments, a comparison of processes should guide the choice of aptitude measures; but the process analysis will also suggest ways in which contrasts between treatments could profitably be formed. Only such analysis will get us out of the present frustrating trial and error.

Although the foregoing paragraphs minimize the relevance of the traditional, very short experiments, often using such artificial stimuli as a paired-associate lists, we do not decry them. Such experiments do advance theory at a relatively basic level, and are suggestive for studies closer to the educational scene. Moreover, they can be conducted in relatively large number. But we have taken very seriously Hawkins' (1966) remark about the "prepared" subject of an experiment; education deals, most of the time, with subjects who are fully habituated -- for better or worse -- to the treatment they are receiving.

Aptitudes

The substantive review of ATI studies in earlier sections has not covered all the studies known to us, and there are surely further studies we have not tracked down. We had originally intended to make our review more comprehensive, and with this end in view a graduate assistant prepared a very large file of abstracts. As has been demon-

strated repeatedly above, the quality of analysis and reporting of research in this field is such, however, that the conclusions of the original author are about as likely to be incorrect as correct. Our assistant was in no position to make the sort of technical critique that a senior member of the staff could, and time has not permitted the senior staff personally to review studies beyond those already discussed. In selecting studies for critical review we have tended to give preference to those found in technical reports and dissertations rather than those in the regular literature, because the former are less likely to be known to the research community and because they generally are reported in more detail and therefore can be better evaluated. Despite the partial nature of our review, we believe that there is sufficient consistency among the studies to warrant general conclusions.

The most basic conclusion from this literature is that simple characterizations of aptitudes and treatments in such terms as "spatial" are unlikely to identify combinations of variables worth investigating. There is no instance where an ATI study defined in terms such as these, using familiar "content" constructs from the PMA, DAT, Guilford or other such aptitude collections, has led to convincing evidence of interaction. The better-controlled studies of this type have led to convincing evidence of absence of interaction. We should not lose sight of the findings of Hills and of Osburn and Melton, which do suggest interactions based on differential aptitudes; but those studies used uncontrolled treatments -- indeed, treatments that were neither described nor reproducible. Consequently, the studies can only suggest a topic for speculation, out of which new studies might come. Some studies that were originally seen as evidence for interactions based on differential abilities, like that of Edgerton, seem on reexamination to be interpretable as interactions resting on general ability.

As to general ability, the first key finding of our survey is that it does seem to be nearly synonymous with "ability to learn", when that term is given its usual commonsense interpretation. There are laboratory tasks to which general ability has little relevance, and there are even some instructional situations employing meaningful content where the correlation of general ability with outcome is slight.

But over and over we have found correlations of broad ability measures or broad composites of verbal reasoning tests with learning outcomes, both in the classroom and under controlled practice conditions. The multiple correlation based on such an ability measure together with a pretest sometimes is as high as 0.70. This is remarkable, since correlations from one learning activity to another are modest, and imply a certain unreliability in the learning process. This confirmation of common sense about the relevance of tested ability is encouraging to the study of ATI, since if one is to have interaction it is necessary to have respectable positive relationship of one aptitude measure with one treatment against which another treatment can be contrasted. The question is how to devise or discover the needed alternative treatment with a flatter regression function.

One strategic line for establishing ATI is to try to develop contrasting treatments, one of which relies heavily on general ability and one of which does not. This statement, like the preceding paragraph, is deliberately vague as to what is meant by "general ability". That term has been used in our summary to capture in one net verbal IQ, nonverbal IQ or more specialized tests of fluid ability such as Hidden Figures, composites of crystallized abilities acquired in school, and collections of cognitive tests in the French or Guilford series. The reason for application of a broad, loose construct is that at present there is no evidence to support a more refined conclusion. It is rare that a study has employed reliable measures of rival subordinate constructs in this domain, and where this has been done one rarely finds that the two kinds of measures give substantially inconsistent predictions of outcome. We shall return to the desirability of sharpening the present overly general construct.

To urge that there must be educationally valuable treatments that do not depend upon the abilities that conventional schooling requires is to carry us straight back to the problem handed to Binet at the start of the century. His test was motivated by the desire to identify, among children performing badly in the Paris schools, the ones who seemed capable of being educated. While he was successful in locating children whose intellectual development was not consistent with their unhappy school records and the unfavorable estimates of teachers,

Binet left the question of what to do with these children to others. His tests and their successors came to be used far more for selection of students who would succeed than for the planning of instruction, because fluid ability predicts success in the established school program less well than crystallized ability, the predictive routine tests used in schools have veered more and more toward measuring crystallized abilities. School psychologists and remedial instructors have had considerable success with underachievers; they have offered special instruction that has advanced the learning of these children. This may or may not be an instance of ATI -- possibly all children would make greater progress if taught by these sensitive, individualized methods. Even if ATI is implied, the greater cost of the individualized instruction seriously restricts its use. There has been no real progress in defining instructional methods, administrable on a large scale, that will improve the learning of pupils who are relatively lacking in aptitude for the regular school program.

There are many indications in our literature review that treatments can be designed that will have this desired effect, i.e., that will yield about the same mean for an unselected group as does the conventional instruction, and that will yield a much smaller regression slope on general ability. We were unable to substantiate the claim that programmed instruction with small steps and overt response would serve this purpose. The contrast of programmed with nonprogrammed instruction, produced no reproducible interactions. Some scattered findings do suggest possibilities of varying (e.g.) the pacing of programmed instruction to produce interactions. These studies could well be repeated and the effects clarified.

The instances that most clearly suggest ATI are diverse, and defy summary. A list of some of the more impressive studies makes clear how diverse the results are:

Kropp et al. find a smaller slope for a "symbolic" presentation of algebra than for a verbal presentation

Edgerton finds a smaller slope for a "meaningful" presentation of technical subject matter

Stallings and Snow find a smaller slope for a "phonics" method than for a "whole word" method in introductory reading

Maier and Jacobs find that conventional instruction has a smaller slope than a combination of teacher and PI
 Herron finds a smaller slope for "new" chemistry instruction than for conventional instruction. (One outcome only).

It does not seem profitable for investigators simply to try haphazard departures from the conventional program in attempting to discover a new hybrid that will serve the student low in conventional aptitudes. This strategy has had dubious results in the past, according to the literature on PI, TV, and homogeneous grouping. What is needed is a theoretical conception of the way in which ability enters into the instructional process.

Theorizing ought to begin by formulating a model of what takes place under conventional instructional techniques. Even the best efforts of this sort to date take us only a short way. Gagné's provocative conception of hierarchies in educational content does much to specify the process of subject-matter learning, but it focusses entirely on content linkages, such as the dependence of two-digit multiplication on the concept of zero as a placeholder. Gagné (and also Ferguson, Fleishman, and others) makes a place for the transfer value of more general aptitudes and integrative processes, but his statements about them are far less articulate than his statements about content. Some thought has been given to processes in laboratory learning, where terms such as "encoding", "short-term memory", and "rehearsal" begin to suggest phenomena to observe and manipulate. So far as educational learning is concerned, the literature we have reviewed provides only one example. The Burton-Goldbeck study separated response learning from associative learning in teaching simple associational material, and found a complex interaction between unfamiliarity of the response to be learned, general ability, and the degree of discrimination required during the instructional treatment. Their theorizing, while worth extending, does not suggest what activities the learner is engaging in that lead to his successes and difficulties under different conditions. As an example of theorizing that comes closer to specifying activities during educational learning we may point to the work of Rothkopf (1965) and others on "mathemagenic behavior". This, however, has not yet been linked to aptitudes.

Perhaps the idea that treatment characteristics are prosthetic devices offers a base for theory. A device compensates for particular aptitude deficiencies in some learners, where other learners can provide needed aptitude function for themselves. For the latter the device may be an overcompensation, interfering with other activities and producing frustration. A treatment without the device allows such learners to capitalize their ability. Perhaps such treatment devices represent externalized mathemagenic behavior, which is either consistent or inconsistent with a given learner's needs. Identifying these consistencies should lead to new aptitude concepts.

So far, then, we have suggested that researchers begin by trying to understand just how the general-ability complex enters into the learning activities of the pupil. Some fraction of that work should be aimed to give a sharp answer as to just what in this ability complex is relevant. The necessary design is to employ at least two reasonably reliable tests of each competing subconstruct, e.g., fluid-analytic ability, ved, perhaps others. The latter is almost certainly the better predictor where new learning depends upon previous lessons, but it is not at all certain which will best predict true residual gains (after taking specifically relevant initial competences --e.g., terminology -- into account).

The second step will be to design alternative treatments on the basis of specific hypotheses about process. The tentative ATI findings to date hint at some possible hypotheses, but as of now these can be only vaguely stated. First, reduction of the burden of semantic processing of verbal information seems likely to give a flat-slope treatment. This can perhaps be done by making the instructional presentation more obvious, through any number of communication devices: easier vocabulary, repetition and paraphrase, sound-tape the learner can hear while he is following the text with his eyes, pictorial elaboration, etc. Some efforts in this direction have failed: we note the Kropp et al. studies of redundancy and the studies of "smooth" programming. But Gagné and Gropper, and Taylor and Fox did have interactions where a pictorial treatment had a flatter slope. Very likely greater success will come when explicit consideration is given to the process the learner goes through. Adding pictures to first readers, we are told, impedes learning to read, even though in one sense it simplifies. In Koran's study

giving more information to student teachers (via video tape) impaired the comprehension of subjects at one ability level and raised it at another.

Second, but closely related, is the possibility of placing on the learner a greater responsibility for organizing material in his own way. There are hints that the abler student responds positively to this kind of challenge, and does less well where his interpretation is constrained by a strong didactic structure. The evidence on this, however, is highly equivocal, and the equivocality leads us to think that the hypothesis is not yet properly stated. The less meaningful, less structured treatment does give scope for the learner to impose an organization -- but it does not generally help the able student. Discovery treatments do not have a reliable advantage for the abler students. The whole concept of autonomous organizing operations by the learner requires sharper definition before theorizing and empirical research can proceed fruitfully. There is the sort of autonomy implicit in self-paced PI and Project PLAN, where the lessons are tightly organized but the learner has considerable choice as to the scheduling of his work; there is the autonomy of instruction that uses unassembled materials such as original documents in history and self-selected problems in mathematics; there is the autonomy that relies on and develops the pupil's self-evaluation rather than on feedback from adults and answer keys; and there is the autonomy of some mathematics projects where the work to be mastered is set, but the learner is encouraged to encode and conceptualize it in his own way so that the answer, when reached, will be "meaningful". So long as all these are indiscriminately referred to as the opposite of didactic or conventional instruction, findings are sure to be confusing.

One who finds a high-slope and a low-slope treatment, considering only general ability or some segment thereof as an aptitude, is in a position to capitalize on ATI in instruction. But if he could find a second aptitude against which the outcome from ^{the} first treatment has a low slope and that from the second a high slope, the allocation decision would yield greater benefits. Moreover, there would be a corner of the bivariate distribution where neither treatment is especially effective, and this would push the search for treatments further. Our next question then, is what sorts of aptitude appear to be candidates for the role of "second", as a foil to "general ability".

One obvious answer is that we should pit the fluid and crystallized segments of general ability against each other. Over the years, the high "nonverbal IQ" has been an embarrassment to educational psychology. Everyone has recognized its reality (not as a fixed quantity, but as a sign that the pupil presently possesses some highly desirable attributes). No one has discovered a generally applicable instructional approach that serves the pupil who is strong on the nonverbal side and relatively weak on the v:ed side. The clinical approach has had its successes, and some cases of this kind have been salvaged by diagnosis of faulty reading and arithmetic skills. But these "under-achievers" have a pervasive difficulty in conventional schoolwork, and there ought to be instructional methods that would serve them better. The candidate methods are precisely those that have the least correlation with v:ed. Hence the second face of the research discussed above is to keep separate measures of v:ed and fluid ability constantly in the picture. If two treatments are found that have similar high and low slopes, respectively, with both kinds of aptitude, there is a useful interaction. If treatments are found where the slope differential is reversed from one aptitude to the other, both practice and theory are beneficiaries. Among the studies we have reviewed, the only solid evidence of such an ATI is the Anderson study, where more meaningful arithmetic profited those who had a record of underachievement, and a more mechanical instruction profited those who had been successful in the past. (We cannot be sure that the same results would emerge today; if it is true that today's conventional instruction is like the rational treatment Anderson considered a novelty in 1940, then the students with adequate achievement today may be those who were underachievers for him.) The Woodruff study also, at least hints at different recommendations for high fluid and high v:ed pupils.

There is another candidate for the spot of "second aptitude". Jensen argues that ability to learn from a strictly rote presentation is distinct from ability to learn through analysis and comprehension, and there is indeed good support for this in studies of the correlates of learning in the laboratory. If deliberate interpretation (e.g., use of mnemonics) cannot turn the task into a meaningful one, then the person with good general ability has no advantage. And there are tests of rote-learning ability ("memory" factors) that do predict success on other such tasks. Jensen calls these "Level I" abilities, and argues that pupils strong on Level I ability should be taught in a

rote manner. Specifically, he argues that lowerclass Negro children will respond well to this instruction despite their difficulties with meaningful lessons.

This raises questions of educational and social policy, as well as of psychological theory and instructional tactics. Unless the rote instruction carries the pupils toward the true educational goals of the common school program, it cannot be defended as a proper application of ATI. To teach some children those skills that can be learned by rote and deprive them of concepts that cannot is not effective education. Even if a great deal of worthwhile content could be taught by capitalizing on Level I abilities, one would have to search for methods of enhancing the analytic level II abilities, so that the pupils high on Level I would not be restricted to rote instruction for all their lives.

In recommending exploitation of the fluid-crystallized distinction, and not of the various "operations", "content", and "product" distinctions handed down via Thurstone, Guilford, and others, we reflect the large amount of negative evidence reviewed. There are no dependable findings of interactions of these hypothesized factors with instructional treatments. We do anticipate that process analysis will lend importance to relatively narrow ability factors or information-processing styles, but we do not find candidate variables among presently prominent "differential" tests.

We undertook our survey of the personality research with considerable hope that this would prove to be a source of useful interactions. The results are almost entirely disappointing. Although hints of interaction can be found, for example in the work on "needs" by McKeachie, Atkinson, and others at the University of Michigan, the effects, weak at best, are inconsistent from one school subject to another and, within an experiment, from one measure of achievement to another. The possible sources of these difficulties are multiple: undependability of any single measure of a personality trait, dependence of behavior on multiple aspects of the personality working simultaneously, difficulty of defining the instructor's style in any simple fashion, etc.

It appears quite premature to set up focussed experiments in which instructional styles are contrasted along a single dimension and an interaction with a single specific personality trait are sought. There is considerable reason to think that the student's personality does

affect his response to the classroom, and there ought to be a steady research effort on this problem. But we cannot recommend the sheer empiricism that in the past has produced dozens of abortive papers on personality scores as predictors of grade average.

The most promising lead is probably the clinical approach of Heil. His results are strong and not implausible. A limited number of classes and teachers was studied, but the investigation was not a small one. There is a need for more studies of precisely this sort, where the investigator becomes intimately familiar with the teachers and their classroom practices. Heil's simple typology should evolve into something more elaborate and better defined, and perhaps it can be linked to some of the variables others have measured in a standard way, such as the teacher's provision of achievement cues. On the aptitude side, we would hope that Heil's clinical methods of classifying pupils could gradually be stated in more rigorous form and based more on reproducible measurements.

Work on personality will have continually to contend with the technical and philosophical problems that arise from the fact that "aptitude" may be a predictor, an intervening variable arising from the treatment and affecting further response to the treatment, or a significant "final" outcome. It will be recalled that Snow and Stallings found possibly important interactions with ability scores as predictors and a "personality" variable -- avoidance of learning activity -- as an outcome variable. This would surely cumulate to disturb development of reading skill if the treatment were long continued, and hence could account for the Grimes-Allinsmith type of interaction with personality. But the personality was one that emerged from the initial reading treatment, rather than something the pupil brought with him to school. We have not discussed the important interactionist studies of Walberg and his colleagues at Harvard Project Physics, since these are for the most part available only in the form of draft manuscript. But they will illustrate the present point very forcibly. A characteristic such as "apathy" is measured by tabulating responses to a questionnaire in which pupils describe the general climate of their classroom; data are collected after the course has been in session for a month or two. Does high apathy imply that the program tended to draw more apathetic students? or that it generated apathy as a local and transient condition, import-

ant only as an intervening variable? or that we have here an early evidence of a permanent detrimental effect on the students' personality, that will radiate into other courses? Obviously, those who study interactions involving personality will have to collect data at several points in time in order to reach an interpretation. This is less critical when ability measures serve as the "aptitude" in an ATI study, just because abilities are less changeable.

Strategy for investigators and supporting agencies

This project was planned to arrive at some conception as to how research on ATI should proceed. While any conception of this sort is open to dispute, it is better to set forth such ideas than to plan solely in terms of the attractiveness of individual research proposals.

Progress toward the goal of identifying and understanding ATI has been slight. We have not examined every pertinent study, but our survey has probed deeply enough to give us confidence that a truly exhaustive sample would not change the general picture as of this moment. There are no solidly established ATI relations even on a laboratory scale and no real sign of any hypothesis ready for application and development. There are intriguing findings here and there, none of which has been pursued through a sufficient series of replication, validity generalization, and enhancement studies to make it impressive.

One reaction to this regrettable stage of affairs would be to abandon ATI research on the grounds that such effects are nonexistent. We urge against this defeatist course. It is inconceivable to us that humans, differing in as many ways as they do, do not differ with respect to the educational treatment that fits each one best. To abandon the ATI model is to assume that there is only one path toward educational development, and that individual differences have no implication save the fatalistic one, of telling the educator that some pupils will advance more rapidly than others no matter what he does.

To argue for a steady effort to detect, define, and ultimately to apply ATI relationships is to encourage high-risk research. The present state of ATI knowledge is reminiscent of that in the simpler area of individual differences at the time Galton's work at South Kensington was in full swing. A generation passed between the initial attention to individual differences that followed on the Origin of Species,

and the arrival in the first decade of the 1900's at socially useful and scientifically important measures. It takes a certain heroism for an investigator or a sponsoring agency to admit that the solution to a problem may well be some distance over the horizon -- but the alternative is the pessimistic one stated in the preceding paragraph.

Research on ATI will have to be subtle. As matters now stand we can be content with neither the conceptualization of aptitude dimensions nor with the conceptualization of treatment dimensions. The former has been studied by excellent investigators for decades, and yet all the questions raised by Spearman and Thurstone are still open. Indeed, the Thurstonian approach which proved extremely useful in the field of vocational assignment (where persons are to do well at different tasks) has proved almost entirely abortive in guiding educational assignment, where we wish to bring persons by different means to master the same tasks. The effort to conceptualize treatment dimensions is almost entirely new, and it is not astonishing that thinking is still at the level of gross concepts such as "difficulty", "degree of structure", and "degree of self-direction". There is reason enough to consider variables of this sort important, despite the ambiguity and need for sharper conceptualization which can only come from continued empirical research combined with thoughtful interpretation.

If it is true that investigators have only rough ideas as to the variables that are most likely to enter into profitable interactions, then research must probe flexibly. We can understand the desire of an agency that supplies funds for research, such as the U. S. Office of Education, to have an explicit plan outlining precisely what will be investigated. But any plan for finding a path through a swamp must be vague, and investigators should be free to deviate with only one constraint -- the requirement that if they strike bottomless ooze, they should warn those who follow behind. To sit on the edge of the swamp and plot out a course saying where the research will be some number of months later is an exercise in naive fantasy. The most profitable research undertakings in this field -- for at least the next decade -- will be those ready to turn on a dime, even if this means abandoning some obsolete "plan" laid down in a research proposal. This characteristic of pioneering research is scarcely unique to the ATI field, but it may be insufficiently appreciated in the Bureau of Educational Research of U.S.O.E.

References

- Aiken, L. R. Interactions among group regressions: An old method in a new setting. Research and Development Memorandum No. 42, Stanford Center for Research and Development in Teaching, Stanford, Calif.: 1968.
- Allison, R. B. Learning parameters and human abilities. Technical Report, Princeton, New Jersey: Educational Testing Service, 1960.
- Alvord, R. W. Learning and transfer in a concept-attainment task: A study in individual differences. Technical Report No. 4. Project on Individual Differences in Learning Ability as a Function of Instructional Variables. Stanford: Stanford University, 1969. Originally a doctoral dissertation, Stanford University, 1967.
- American Institutes for Research. Annual Report, Pittsburgh, Pa. American Institutes for Research, 1965, pp. 15-16.
- Anderson, G. L. A comparison of the outcomes of instruction under two theories of learning. Unpublished doctoral dissertation, University of Minnesota, 1941. See also in E. J. Swenson, et al. (Eds.) Learning theory in school situations. Minneapolis: University of Minnesota Press, 1949.
- Anderson, J. E. The limitations of infant and pre-school tests in the measurement of intelligence. Journal of Psychology, 1939, 8, 351-379.
- Atkinson, J. W., & Feather, N. (Eds.) A theory of achievement motivation. New York: Wiley, 1966.
- Becker, J. P. An attempt to design instructional techniques in mathematics to accomodate different patterns on mental ability. Unpublished doctoral dissertation, Stanford University, 1967.
- Behr, M. J. Interactions between "structure-of-intellect" factors and two methods of presenting concepts of modulus seven arithmetic. Unpublished doctoral dissertation, Florida State University, 1967.
- Blatt, S. J. Effects of test anxiety and instructional context on problem solving. Cooperative Research Project, 1382, 1963.
- Bloom, B. S. Stability and change in human characteristics. New York: Wiley, 1964.
- Bogarty, R. S. The criterion method: Some analyses and remarks. Psychological Bulletin, 1965, 64, 1-14.
- Bond, G. L. The auditory and speech characteristics of poor readers. Bureau of Publication, Teachers College, Columbia University, 1935.

- Bottenberg, R. A., & Ward, J. H., Jr. Applied multiple linear regression. (PRL-TDR-63-6) Lackland Air Force Base, Texas, 1963.
- Briggs, L. J. Sequencing of instruction in relation to comments. Pittsburgh, Pa.: American Institutes for Research, 1968.
- Brogden, H. E. Increased efficiency of selection resulting from replacement of a single predictor with several differential predictors. Educational and Psychological Measurement, 1951, 11, 173-196.
- Brownell, W. A., & Moser, A. G. Meaningful versus mechanical learning: A study in grade three subtraction. Duke University Research Studies in Education, No. 8. Durham, N. C.: Duke University Press, 1949.
- Bunch, M. E. The amount of transfer in rational learning as a function of time. Journal of Comparative Psychology, 1936, 22, 325-337.
- Bunch, M. E., & McCraven, V. G. The temporal course of transfer in the learning of memory material. Journal of Comparative Psychology, 1938, 25, 481-496.
- Bunderson, C. V. Transfer of mental abilities at different stages of practice in the solution of concept problems. Research Bulletin, RB-67-20, Educational Testing Service, and Office of Naval Research Technical Report, 1967. Originally a doctoral dissertation, Princeton University, 1965.
- Burket, G. R. A study of reduced rank models for multiple prediction. Psychometric Monographs, 1964, No. 12.
- Burton, B. B., & Goldbeck, R. A. The effect of response characteristics in multiple life and choice alternatives on learning during programed instruction. San Mateo, Calif.: American Institutes for Research, 1962.
- Bush, R. R., & Lovejoy, E. P. Learning to criterion: A study of individual differences. Stanford Colloquium, April 28, 1965. Mimeo.
- Bush, W. J., Gregg, D. K., Smith, E. A., & McBride, C. B. Some interactions between the individual differences and modes of instruction. USAF AMRL Technical Report No. 65-228, iii, 1965.
- Butler, W. I. A college English teacher looks at television: Composition. Journal of Educational Sociology, 1958, 31, 346-352.
- Cahoon, D. D., Peterson, L. P., & Watson, C. G. Relative effectiveness of programed text and teaching machine as a function of measured interest. Journal of Applied Psychology, 1968, 52, 454-456.

- Campbel, V. N. Bypassing as a way of adapting instructional programs to individual differences. Journal of Educational Psychology, 1963, 54, 337-345.
- Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. In Gage, N. L. (Ed.) Handbook of research on teaching. Chicago: Rand McNally, 1963.
- Campeau, P. Level of anxiety and presence on absence of feedback in PI. Palo Alto, Calif.: Final Report 7-28-7670-229, American Institutes for Research, 1965.
- Campeau, P. L. Selective review of literature on audiovisual media of instruction. In L. J. Briggs, et al. Instructional media. Palo Alto: American Institutes for Research, 1967. Pp. 99-142.
- Carroll, J. B. The prediction of success in intensive foreign language training. In R. Glaser (Ed.), Training research and education. Pittsburgh: University of Pittsburgh Press, 1962. Pp. 87-136.
- Carroll, J. B. A model of school learning. Teachers College Record, 1963, 64, 723-733.
- Carroll, J. B. School learning over the long haul. In J. D. Krumboltz (Ed.), Learning and the educational process. Chicago: Rand McNally, 1965.
- Carroll, J. B., & Leonard, G. The effectiveness of programmed "Grafdriels" in teaching the Arabic writing system. Cambridge: Laboratory for Research in Instruction, Graduate School of Education, Harvard University, 1963.
- Carry, L. R. Interaction of visualization and general reasoning abilities in curriculum treatment in algebra. Unpublished doctoral dissertation, Stanford University, 1967.
- Cartwright, G. P. Two types of programed instruction for mentally retarded adolescents. Unpublished Masters' thesis, University of Illinois, 1962. Summarized by Stolurow (1964).
- Cohen, J. Multiple regression as a general data-analytic system. Psychological Bulletin, 1968, 70(6), 426-443.
- Corsini, D. A., Pick, A. D., & Flavell, J. H. Production deficiency of non-verbal mediators in young children. Child Development, 1968, 39, 53-58.
- Corwin, E. S. The impact of the idea of evolution on the American political and constitutional tradition. In Pearson, S. (Ed.) Evolutionary thought in America. Yale University Press, 1950, Pp. 182-199.

- Covington, M. The effect of anxiety on various types of ideational output measures in complex problem-solving. Paper read at annual Western Psychological Association convention, San Francisco, 1967.
- Cronbach, L. J. Correlation between persons as a research tool. In O. H. Mowrer (Ed.) Psychotherapy: Theory and problems. New York: Ronald, 1953. Pp. 376-388.
- Cronbach, L. J. Processes affecting scores on "understanding of others" and "assumed similarity." Psychological Bulletin, 1955, 52, 177-193.
- Cronbach, L. J. The two disciplines of scientific psychology. American Psychologist, 1957, 12, 671-684.
- Cronbach, L. J. Educational psychology. (2nd ed.) New York: Harcourt, Brace, & World, 1963.
- Cronbach, L. J. Psychological background for curriculum experimentation. In P. C. Rosenbloom and P. C. Hillestad (Eds.) Modern viewpoints in curriculum. New York: McGraw Hill, 1964.
- Cronbach, L. J. Year-to-year correlations of mental tests: A review of the Hofstetter analysis. Child Development, 1967, 38, 283-289.
- Cronbach, L. J. How can instruction be adapted to individual differences? In R. M. Gagné (Ed.) Learning and individual differences. Columbus, Ohio: Charles E. Merrill Books, 1967. Pp. 23-39.
- Cronbach, L. J. Intelligence? Creativity? A parsimonious reinterpretation of the Wallach-Kogan data. Technical report No. 2. Project on Individual Differences in Learning Ability as a Function of Instructional Variables, Stanford University, 1968. American Educational Research Journal, 1968, 5, 491-512.
- Cronbach, L. J. Heredity, environment, and educational policy. Harvard Educational Review, Spring, 1969, in press.
- Cronbach, L. J., & Furby, L. How we should measure "change" - or should we? Technical Report No. 6, Project on Individual Differences in Learning Ability as a Function of Instructional Variables. Stanford: Stanford University, 1969.
- Cronbach, L. J., & Gleser, G. C. Psychological tests and personnel decisions. (2nd ed.) Urbana: University of Illinois Press, 1965.
- Cronbach, L. J., & Suppes, P. C. (Eds.) Disciplined inquiry for education. New York: Macmillan, in press, 1969.

- Curry, R. P. Report of three experiments on the use of television in instruction. Cincinnati, Ohio: Cincinnati Public Schools, 1959.
- Curry, R. P. Report of four experiments in the use of television in instruction. Cincinnati, Ohio: Cincinnati Public Schools, 1960.
- Della-Piana, G. An experimental evaluation of programmed learning. Journal of Educational Research, 1962, 55, 495-501.
- Domino, G. Differential predictions of academic achievement in conforming and independent settings. Journal of Educational Psychology, 1968, 59, 256-260.
- Doty, B. A., & Doty, L. A. Programed instructional effectiveness in relation to certain student characteristics. Journal of Educational Psychology, 1964, 55, 334-338.
- Dreyer, R. E., & Beatty, W. H. Instructional television research. Project #1: An experimental study of college instruction using broadcast television. San Francisco: San Francisco State College, 1958.
- DuBois, P. H. Multivariate correlational analysis. New York: Harper, 1957.
- Duncanson, J. P. Intelligence and the ability to learn. Princeton, N. J.: Educational Testing Service Research Bulletin (RB-64-29), and Office of Naval Research Technical Report, 1964. (See also Learning and measured abilities. Journal of Educational Psychology, 1966, 57, 220-229.
- Dunham, J. L., & Bunderson, V. The effect of rule instruction upon the relationship of cognitive abilities to performance in multiple-category concept learning problems. Paper presented to American Educational Research Association, 1968.
- Dunham, J. L., Guilford, J. P., Hoepfner, R. Abilities pertaining to classes and the learning of concepts. Reports, Psychological Laboratory, University of Southern California, No. 39, 1966. (See also Psychological Review, 1968, 75, 206-221, for a shorter account).
- Dunlap, K. A revision of the fundamental law of habit formation. Science, 1928, 67, 360-362.
- Edgerton, H. A. The relationship of method of instruction to trainee aptitude pattern. Technical Report, Contract Nonr 1042 (00). New York: Richardson, Bellows, Henry & Co., 1958.
- Eimas, P. D. Effects of overtraining and age on intradimensional and extra-dimensional shifts in children. Journal of Experimental Child Psychology, 1966, 3, 348-355.

- Endler, N. S., Hunt, J. McV., & Rosenstein, A. J. An S-R inventory of anxiousness. Psychological Monographs, 1962, 76, No. 17.
- Eysenck, H. J. Dynamics of anxiety and hysteria. London: Routledge & Kegan Paul, 1957.
- Eysenck, H. J. Intelligence assessment: A theoretical and experimental approach. British Journal of Educational Psychology, 1967, 37, 81-98.
- Feldman, M. E. Learning by programmed and text format at three levels of difficulty. Journal of Educational Psychology, 1965, 56, 133-139.
- Ferguson, G. A. On learning and human ability. Canadian Journal of Psychology, 1954, 8, 95-112.
- Ferguson, G. A. On transfer and the abilities of man. Canadian Journal of Psychology, 1956, 10, 121-131.
- Ferguson, L. R., & Maccoby, E. Interpersonal correlates of differential abilities. Child Development, 1966, 37, 549-571.
- Ferris, F. L., Jr. Testing in the new curriculum, neuromorology, "tyranny", or common sense? School Review, 1962, 70, 112-131.
- Fitzgerald, D., & Ausubel, D. P. Cognitive versus affective factors in the learning and retention of controversial material. Journal of Educational Psychology, 1963, 54, 73-84.
- Flanders, N. A. Teacher influence, pupil attitudes, and achievement. Cooperative Research Monograph No. 12. Ann Arbor, Mich.: University of Michigan, 1965.
- Flavell, J. H., Beach, D. H., & Chinsky, J. M. Spontaneous verbal rehearsal in the memory task as a function of age. Child Development, 1966, 37, 283-299.
- Fleishman, E. A. Human abilities and the acquisition of skill. In E. A. Bilodeau (Ed.) Acquisition of skill. New York: Academic Press, 1966.
- Fleishman, E. A., & Hempel, W. E., Jr. Changes in factor structure of a complex psychomotor test as a function of practice. Psychometrika, 1954, 19, 239-252.
- Freibergs, V., & Tulving, E. The effect of practice on utilization of information from positive and negative instances in concept identification. Canadian Journal of Psychology, 1961, 15, 101-106.
- Frederiksen, C. H. Abilities, transfer, and information retrieval in verbal learning. Technical Report, Contract Nonr 1834 (9), University of Illinois, Department of Psychology. 1967.

- Gagné, R. M. Ability differences in the learning of concepts governing directed numbers. In Research problems in mathematics education. Cooperative Research Monographs, No. 3, 1960, 112-113.
- Gagné, R. M. (Ed.) Learning and individual differences. Columbus, Ohio: Charles E. Merrill Books, 1967.
- Gagné, R. M., & Gropper, G. L. Individual differences in learning from visual and verbal presentations. Pittsburgh, Pa.: American Institutes for Research, 1965.
- Gagné, R. M., & Paradise, N. E. Abilities and learning sets in knowledge acquisition. Psychological Monographs, 1961, 75, No. 14.
- Gifford, E., & Marston, A. R. Test anxiety, reading rate and task experience. Journal of Educational Research, 1966, 59, 303-310.
- Greenbaum, C. W. Effect of situational and personality variables on improvisation and attitude change. Journal of Personality and Social Psychology, 1966,
- Grimes, J. W., & AllinSmith, A. W. Compulsivity, anxiety, and school achievement. Merrill-Palmer Quarterly, 1961, 7, 247-271. Reprinted in Rosenblith, J. F., & AllinSmith, A. W. (Eds), The causes of behavior. Allyn & Bacon, 1966.
- Gropper, G. L. Controlling student response during visual presentations. Pittsburgh, Pa.: American Institutes for Research, 1965.
- Guilford, J. P. The nature of human intelligence. New York: McGraw Hill, 1967.
- Guilford, J. P., Hoepfner, R., & Peterson, H. Predicting achievement in ninth-grade mathematics from measures of intellectual aptitude factors. Educational and Psychological Measurement, 1965, 25, 659-682.
- Gulliksen, H. Louis Leon Thurstone, experimental and mathematical psychologist. American Psychologist, 1968, 23, 786-802.
- Guttman, L. The structure of relations among intelligence tests. Proceedings, 1964 Invitational Conference on Testing Problems. Princeton, N. J. Educational Testing Service, 1965. Pp. 25-36.
- Guttman, L. Order analysis of correlation matrices. In Cattell, R. B. (Ed.) Handbook of multivariate experimental psychology. Chicago: Rand McNally, 1966.
- Guttman, L., & Schlesinger, I. M. The analysis of diagnostic effectiveness of a facet design battery of achievement and analytical ability tests. Jerusalem, Israel Institute of Applied Social Research, 1967.

- Hamilton, N. R. Differential response to instruction designed to call upon spatial and verbal aptitudes. Technical Report No. 5. Project on Individual Differences in Learning Ability as a Function of Instructional Variables. Stanford: Stanford University, 1969. Originally a doctoral dissertation, Stanford University, 1968.
- Harlow, H. F. The formation of learning sets. Psychological Review, 1949, 56, 51-56.
- Harlow, H. F. Learning set and error factor theory. In S. Koch (Ed.), Psychology: A study of a science. Vol. 2 McGraw-Hill, 1959. Pp. 492-537.
- Harris, C. W. (Ed.), Problems in measuring change. Madison, Wisc.: University of Wisconsin Press, 1963.
- Harris, C. W. On factors and factor scores. Psychometrika, 1967, 32, 363-379.
- Hawkins, D. Learning the unteachable. In L. Shulman and E. Keislar, (Eds.), Learning by discovery: A critical appraisal. Chicago: Rand McNally, 1966. Pp. 3-12.
- Heckhausen, H. The anatomy of achievement motivation. New York: Academic Press, 1967.
- Heil, L. M., and others. Characteristics of teacher behavior related to the achievement of children in several elementary grades. Cooperative Research Project No. 352. Mimeo. Brooklyn: Brooklyn College, 1960.
- Herron, J. D. Evaluation of the new curricula. Journal of Research on Science Teaching, 1966, 4, 159-170.
- Hershberger, W. Self-evaluational responding and typographical cueing. Journal of Educational Psychology, 1964, 55, 288-296.
- Hills, J. R. Factor-analyzed abilities and success in college mathematics. Educational and Psychological Measurement, 1957, 17, 615-622.
- Hoepfner, R., Guilford, J. P., & Merrifield, P. R. A factor analysis of the symbolic-evaluation abilities. Psychological Laboratory, University of Southern California, Reports, 1964, No. 33.
- Hoepfner, R., & Guilford, J. P. Figural, symbolic, and semantic factors of creative potential in ninth-grade students. Psychological Laboratory, University of Southern California, Reports, 1965, No. 35.

- Hofstaetter, P. R. The changing composition of "intelligence": A study in T-technique. Journal of Genetic Psychology, 1954, 85, 159-164.
- Holland, J. L. The psychology of vocational choice. Waltham, Mass.: Blaisdell, 1966.
- Horn, J., & Cattell, R. B. Refinement and test of the theory of fluid and crystallized intelligence. Journal of Educational Psychology, 1966, 57, 253-276.
- Husband, R. W. Positive transfer as a factor in memory. Proceedings of Iowa Academy of Sciences, 1947, 54, 235-238.
- Hutchinson, W. L. Creative and productive thinking in the classroom. Unpublished doctoral dissertation, University of Utah, 1963.
- Iwahara, S., & Tanabe, N. Anxiety level and the effect of verbal reinforcement combinations on a verbal-coding task. Japanese Psychological Research, 1963, 5, 147-152.
- Jacobs, J. N., & Bollenbacher, J. K. An experimental study of the effectiveness of television vs. classroom instruction in sixth-grade science in the Cincinnati Public Schools, 1956-1957. Journal of Educational Research, 1959, 52, 184-189.
- Jacobs, J. N., Bollenbacher, J. K., & Keiffer, M. Teaching seventh-grade mathematics by television homogeneously grouped below average students. The Mathematics Teacher, 1961, 54, 551-555.
- Jacobs, P. I., Maier, M. H., & Stolurow, L. M. A guide to evaluating self-instructional programs. New York: Holt, Rinehart, & Winston, 1966.
- Jensen, A. R. How much can we boost IQ and scholastic achievement? Harvard Educational Review, 1969, TK.
- Jensen, A. , & Rohwer, W. D., Jr. Syntactical mediation of serial and paired associative learning as a function of age. Child Development, 1965, 36, 601-608.
- Johnson, P. O., & Neyman, J. Tests of certain linear hypotheses and their applications to some educational problems. Statistical Research Memoirs, 1936, 1, 57-63.
- Kagan, J. Learning, attention and the issue of discovery. In L. S. Shulman & E. R. Keislar (Eds.), Learning by discovery: A critical appraisal. Chicago: Rand McNally, 1966.

- Kagan, J., Rosman, B. L., Day, D., Albert, J., & Philips, W. Information processing in the child: Significance of analytic and reflective attitudes. Psychological Monographs, 1964, 78, No. 1. (Whole No. 578)
- Katahn, M. Interaction of anxiety and ability in complex learning situations. Journal of Personal and Social Psychology, 1966, 3, 475-479.
- Katahn, M., & Lyda, L. L. Anxiety and the learning of responses varying in initial rank in the response hierarchy. Journal of Personality, 1966, 34, 287-299.
- Keeney, T. J., Cannizzo, S. R., & Flavell, J. H. Spontaneous and induced verbal rehearsal in a recall task. Child Development, 1967, 38, 953-966.
- Kight, H., & Sassenrath, J. M. Relation of achievement motivation and test anxiety to performance in programmed instruction. Journal of Educational Psychology, 1966, 57, 14-17.
- Konstadt, N., & Forman, E. Field dependence and external directedness. Journal of Personal and Social Psychology, 1965, 1, 490-493.
- Koran, M. L. The effect of individual differences on observational learning in the acquisition of a teaching skill. Unpublished doctoral dissertation, Stanford University, 1969.
- Korchin, S. J., & Levine, S. Anxiety and verbal learning. Journal of Abnormal and Social Psychology, 1957, 54, 234-240.
- Kress, G. C., Jr., & Gropper, G. L. A comparison of two strategies for individualizing fixed-paced programmed instruction. American Educational Research Journal, 1966, 3, 273-280.
- Kropp, R. P., Nelson, W. H., & King, F. J. Identification and definition of subject-matter content variables related to human aptitudes. Unpublished report, Cooperative Research Project No. 2914. Tallahassee, Florida: Florida State University, 1967.
- Kruskal, J. B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika, 1964, 29, 1-27.
- Kruskal, J. B. Nonmetric multidimensional scaling: A numerical method. Psychometrika, 1964, 29, 28-42.
- Leith, G. O., & Bossett, R. Mode of learning and personality. Research report on programmed learning. No. 14. Mimeo. University of Birmingham School of Education, 1967.
- Lawrence, D. H. The evaluation of training and transfer programs in terms of efficiency measures. Journal of Psychology, 1954, 38, 367-382.

- Levin, G. R., & Baker, B. C. Item scrambling in a self-instructional program. Journal of Educational Psychology, 1963, 54, 138-143.
- Li, J. C. R. Statistical inference. Vol. 2. The multiple regression and its ramifications. Ann Arbor, Mich.: Edwards Bros., 1964.
- Lim, K. B. Prompting vs. confirmation, pictures vs. translations, and other variables in children's learning of grammar in a second language. Unpublished doctoral dissertation, Harvard University, 1968.
- Lingoes, J. C. An IBM 0790 program for Guttman-Lingoes smallest space analysis-I. Behavioral Science, 1965, 10, 183-184.
- Lord, F. M. Elementary models for measuring change. In C. W. Harris (Ed.) Problems in measuring change. Madison, Wisc.: University of Wisconsin Press, 1963.
- Lublin, S. C. Reinforcement schedules, scholastic aptitude, autonomy need, and achievement in a programmed course. Journal of Educational Psychology, 1965, 56, 295-302.
- Maccoby, E. The development of sex differences. Stanford: Stanford University Press, 1966.
- Madansky, A. The fitting of straight lines when both variables are subject to error. Journal of the American Statistical Association, 1959, 54, 173-205.
- Maier, M. H., & Jacobs, P. I. Programed learning -- some recommendations and results. Bulletin of the National Association of Secondary School Principals, 1964, 48, 242-255.
- Maier, M., & Jacobs, P. B. The effects of variations in a self-instructional program on instructional outcomes. Psychological Reports, 1966, 18, 539-546.
- Mandler, G., & Sarason, S. B. A study of anxiety and learning. Journal of Abnormal and Social Psychology, 1952, 47, 166-173.
- Manley, M. B. A factor analytic study of three types of concept attainment tasks. Princeton, N. J.: Educational Testing Service Research Bulletin (RB-65-31), October 1965.
- McCandless, B. R., & Castenada, A. Anxiety in children, school intelligence, and achievement. Child Development, 1956, 27, 379-382.
- McGeogh, J. A., & Irion, A. L. The psychology of human learning. Longmans Green, 1952.
- McKeachie, W. J. Students, groups, and teaching methods. American Psychologist, 1958, 13, 580-584.

- McKeachie, W. J. Motivation, teaching methods, and college learning. In M. R. Jones (Ed.), Nebraska Symposium on motivation, 1961. Lincoln, Neb.: University of Nebraska, 1961. Pp. 111-142.
- McKeachie, W. J., and others. Student affiliation motives, teacher warmth, and academic achievement. Journal of Personality and Social Psychology, 1966, 4, 457-461.
- McLaughlin, R. J., & Eysenck, H. J. Extroversion, neuroticism, and paired-associates learning. Journal of Experimental Research in Personality, 1967, 2, 128-132.
- McNeil, J. D. Programed instruction vs. usual classroom procedures in teaching boys to read. American Educational Research Journal, 1964, 1, 113-119.
- McNeil, J. D., & Keislar, E. R. Value of the oral response in beginning reading: An experimental study using PI. British Journal of Educational Psychology, 1963, 33, 162-168.
- McNemar, Q. On growth measurement. Educational and Psychological Measurement, 1958, 18, 47-55.
- Melton, A. W. Individual differences and theoretical process variables: general comments on the conference. In R. M. Gagné, (Ed.) Learning and individual differences. Columbus, Ohio: Charles E. Merrill Books, 1967. Pp. 238-252.
- Merrill, D. D., & Stolurow, M. Hierarchical preview vs. problem-oriented review in learning an imaginary science. American Educational Research Journal, 1966, 3, 251-261.
- Mischel, W. Personality and assessment. New York: Wiley, 1968.
- Moely, B. E., Olson, F. A., Halwes, T.G., & Flavell, J. H. Production deficiency in young children's clustered recall. Developmental Psychology, 1969, 1, 26-34.
- Nanda, H. Factor analytic techniques for inter-battery comparison and their application to some psychometric problems. Unpublished doctoral dissertation, Stanford University, 1967.
- Neale, J. M., & Katahn, M. Anxiety, choice and stimulus uncertainty. Journal of Personality, 1968, 36, 235-245.
- O'Connor, P., Atkinson, J. W., & Horner, M. Motivational implications of ability grouping in school. In J. W. Atkinson & N. T. Feather, (Eds.) A theory of achievement motivation. New York: Wiley, 1966.

- Osburn, H. G., & Melton, R. F. Prediction of proficiency in a modern and traditional course in beginning algebra. Educational and Psychological Measurement, 1963, 23, 277-288.
- Osler, S. F. The cognitive status of the disadvantaged children. Paper read at the American Psychological Association Annual Convention, San Francisco, September, 1968.
- Pace, C. R., & Stern, G. G. An approach to the measurement of physiological characteristics of college environments. Journal of Educational Psychology, 1958, 49, 269-277.
- Patton, J. A. A study of the effects of student acceptance of responsibility and motivation on course behavior. Unpublished doctoral dissertation, University of Michigan, 1955.
- Payne, D. A., Krathwohl, D. R., & Gordon, J. The effect of sequence in programmed instruction. American Educational Research Journal, 1967, 4, 123-132.
- Peterson, D. R. The clinical study of social behavior. New York: Appleton-Century-Crofts, 1968.
- Reed, J. E., & Hayman, J. L., Jr. An experiment involving use of English 2600, an automated instructional text. Journal of Educational Research, 1962, 55, 476-484.
- Reese, H. W. Discrimination learning set in children. In L. Lipsitt & C. Spiker (Eds.), Advances in child development and behavior. Vol. I. New York: Academic Press, 1964, Pp. 116-145.
- Rin, Y. Extroversion, neuroticism and the effect of praise or blame. British Journal of Educational Psychology, 1965.
- Ripple, R., & O'Reilly, G. Abstract. Research in Education, September, 1967, No. 9, ED 011 072.
- Rosenthal, R., & Jacobson, L. Pygmalion in the classroom: Teacher expectation and pupils' intellectual development. New York: Holt, Rinehart, Winston, 1968.
- Rothkopf, E. Z. Some theoretical and experimental approaches to problems in written instruction. In J. D. Krumboltz, (Ed.) Learning and the educational process. Chicago: Rand McNally, 1965, Pp. 193-221.
- Ryan, F. L. Advance organizers and test anxiety in programmed social studies instruction. California Journal of Educational Research, 1968, 19, 67-76.
- Ryan, W. D., & Lakie, W. L. Competitive and noncompetitive performance in relation to achievement motive and manifest anxiety. Journal of Personality and Social Psychology, 1965, 1, 342-345.

- Salomon, G. Interaction of communication-medium and two procedures of training for subjective response uncertainty of teachers. Unpublished doctoral dissertation, Stanford University, 1968.
- Scharf, E. S. A study of the effects of partial reinforcement on a behavior in a programmed learning situation. Chapter 4 in R. Glaser & J. I. Taber (Eds.) Investigations of the characteristics of programmed learning sequences. Pittsburgh: University of Pittsburgh, Dept. of Psychology, 1961.
- Scheffé, H. The analysis of variance. New York: Wiley, 1959.
- Seibert, W. F., & Snow, R. E. Studies in cine-psychometry I: preliminary factor analysis of visual cognition and memory. Final Report to USOE. Grant No. 7-12-0280-184, Audio Visual Center, Purdue University, Lafayette, 1965.
- Seibert, W. F., Reid, J. C., & Snow, R. E. Studies in cine-psychometry II: Continued factoring of audio and visual cognition and memory. Final Report to USOE. Grant No. 7-24-0280-257, Audio Visual Center, Purdue University, Lafayette, 1967.
- Shepard, R. N. The analysis of proximities: Multidimensional scaling with an unknown distance function. Psychometrika, 1962, 27, 125-139; 219-246.
- Shepard, R. N. A simplicial design for the analysis of correlational learning data. Multivariate Behavioral Research, 1967, 83-87.
- Shepard, R. N., & Carroll, J. D. Parametric representation of nonlinear data structures. In P. R. Krishniah, (Ed.) Multivariate analysis. Academic Press, 1966, Pp. 561-592.
- Silberman, H. et al. Development and evaluation of self-instructional materials for underachieving and overachieving students. Santa Monica: Technical Memorandum 727, System Development Corporation, 1962.
- Skinner, B. F. The science of learning and the art of teaching. Harvard Educational Review, 1954, 25, 86-97.
- Skinner, B. F. Teaching science in high school -- what is wrong? Science, 1968, 159, 704-710.
- Smith, L.M. Programed learning in elementary school: An experimental study of relationships between mental abilities and performance. Technical Report, USOE Title VII Project 71151.01. University of Illinois, Training Research Laboratory, 1962. Also an unpublished doctoral dissertation.

- Smock, C. D. Perceptual rigidity and closure phenomenon as a function of manifest anxiety in children. Child Development, 1958, 29,
- Snow, R. E. Review of Rosenthal, R., & Jacobson, L., Pygmalion in the classroom: Teacher expectation and pupils' intellectual development. Contemporary Psychology, 1969, in press.
- Snow, R. E., & Salomon, G. Aptitudes and instructional media. Technical Report, No. 3. Project on Individual Differences in Learning Ability as a Function of Instructional Variables. Stanford, Calif: Stanford University, 1968. See also AV Communication Review, 1968, 16, 341-357.
- Snow, R. E., Tiffin, J., & Seibert, W. F. Individual differences and instructional film effects. Journal of Educational Psychology, 1965, 56, 315-326.
- Spence, K. W. The reliability of the maze and methods of its determination. Comparative Psychological Monographs, 1932, 8, 1-45.
- Stanley, J. C. On improving certain aspects of educational experimentation. In Improving experimental design in statistical analysis. Chicago: Rand, McNally, 1967. Pp. 1-27.
- Stevenson, H. W., & Odom, R. D. Interrelationships in children's learning. Child Development, 1965, 36, 7-19.
- Stevenson, H. W., Hale, G. A., Klein, R. E., & Miller, L. K. Interrelations and correlates in children's learning and problem solving. Mimeographed, 1968.
- Stolurow, L. M. Social impact of programmed instruction: Aptitudes and abilities revisited. In J. P. DeCecco, (Ed.) Educational technology. New York: Holt, Rinehart & Winston, 1964. Pp. 348-355.
- Stolurow, L. M. Programmed instruction and teaching machines. In P. H. Rossi & B. J. Biddle, (Eds.) The new media and education. Chicago: Aldine, 1966. Pp. 124-177.
- Tallmadge, G. K., Shearer, J. W., & Greenberg, A. Study of training equipment and individual differences: The effects of subject matter variances. Technical Report: NAVTRADEVCEEN 67-C-0114-1, American Institutes for Research, Palo Alto, Calif., May 1968.
- Tanner, R. T. Expository-deductive vs. discovery-inductive programing of physical science principles. Unpublished doctoral dissertation, Stanford University, 1968.

- Taylor, J. E., & Fox, W. L. Differential approaches to training. Professional paper 47-67. Alexandria, Va.: Human Resources Research Office, 1967.
- Thelen, H. A. Classroom grouping for teachability. New York: Wiley, 1967.
- Thorndike, R. L. Organization of behavior in the albino rat. Genetic Psychological Monographs, 1935, 17, No. 1.
- Tolman, E. C., & Nyswander, D. B. The reliability and validity of maze measures for rats. Journal of Comparative Psychology, 1927, 7, 425-260.
- Traub, R. E. The importance of problem heterogeneity to programmed learning. Doctoral dissertation, Princeton University, 1964. Issued as RB 64-26, Educational Testing Service, 1964. Journal of Educational Psychology, 1966, 57, 54-60.
- Tucker, L. R. An inter-battery method of factor analysis. Psychometrika, 1958, 23, 111-136.
- Tucker, L. R., Damarin, F., & Messick, S. A base-free measure of change. Psychometrika, 1966, 31, 457-473.
- Vernon, P. E. The structure of human abilities. London: Methuen, 1961.
- Vernon, P. E. Ability factors and environmental influences. American Psychologist, 1965, 20, 723-733.
- Wallach, M. A., & Kogan, N. Modes of thinking in young children. New York: Holt, Rinehart, & Winston, 1965.
- Wicklegren, W., & Cohen, D. H. An artificial language and memory approach to concept attainment. Psychological Reports, 1962, 11, 815-827.
- Williams, J. Comparison of several response modes in a review program. Journal of Educational Psychology, 1963, 54, 253-260.
- Williams, J. Effectiveness of constructed response and multiple-choice programming modes as a function of test mode. Journal of Educational Psychology, 1965, 45, 111-117.
- Williams, J. P., & Levy, E. I. Retention of introductory and review programs as a function of response mode. American Educational Research Journal, 1964, 1, 211-218.
- Witkin, H. A. et al. Psychological differentiation. New York: Wiley, 1962.
- Wittrock, W. C. Response mode in the programming of kinetic molecular-theory concepts. Journal of Educational Psychology, 1963, 54, 89-93.

- Wolff, J. L. Effects of verbalization and pretraining on concept attainment by children in two mediation categories. Journal of Educational Psychology, 1967, 58, 19-27.
- Woodrow, H. The ability to learn. Psychological Reviews, 1946, 53, 147-158.
- Woodruff, A. B., Shimabukuro, S., & Frey, S. H. Methods of programmed instruction related to student characteristics. DeKalb, Illinois: Northern Illinois University, 1965.
- Yerkes, R. M., & Dodson, J. D. The relation of strength of stimulus to rapidity of habit-formation. Journal of Comparative Neurological Psychology, 1908, 18, 459-482.
- Zeaman, D., & House, B. J. The relation of IQ and learning. In R. M. Gagné (Ed.) Learning and individual differences. Columbus, Ohio: Merrill, 1967, Pp. 192-212.